

6.2: Ratios and probabilities

Introduction

Let's define our terms. An **event** is some occurrence. As you know, a ratio is one number, the numerator, divided by another, called the denominator. A **proportion** is a ratio where the numerator is a part of the whole. A **rate** is a ratio of the frequency of an event during a certain period of time. A rate may or may not be a proportion, and a ratio need not be a proportion, but proportions and rates are all kinds of ratios. If we combine ratios, proportions, and/or rates, we construct an **index**.

Ratios

Yes, data analysis can be complicated, but we start with this basic idea. Much of the statistics is based on frequency measures, e.g., ratios, rates, proportions, indexes, and scales.

Ratios are the association between two numbers, one random variable divided by another. Ratios are used as descriptors and the numerator and denominator do not need to be of the same kind. Business and economics are full of ratios. For example, Return on investment (ROI) equals net income divided by number of shares outstanding, the Price-Earnings, or P/E ratio, is the ratio of the price of a stock to the earnings per stock, as well as many others are used to summarize performance of a business, and to compare performance of one business against another. Ratios are a deceptively convenient way to standardize a variable for comparisons, i.e., how many times one number contains another. For example, when estimating bird counts for different areas, or different birding effort (intensity, time searched), we may correct counts by accounting for area in which counts were made or the total time spent counting, for a per-unit ratio (Liermann et al 2004).

Practice: There were 1,326 day undergraduate students enrolled in 2014 at Chaminade University of Honolulu and the Sullivan Library added 8469 new items (ebooks, journals, etc.,) to its collection during 2014. What is the ratio of new items per student?

$$\frac{8469 \text{ items}}{1326 \text{ students}} = 6.39 \text{ items per student}$$

Data collected from Chaminade University website at www.chaminade.edu on 3 July 2014.

Practice: For another example, what is ratio of annual institutional aid a student at Chaminade University may expect to receive compared to a student at Hawaii Pacific University?

$$\frac{\text{aid}_{\text{Chaminade}}}{\text{aid}_{\text{HPU}}} = \frac{\$8491}{\$3897} = 2.3$$

Fold-change

To compare the ratio between two quantities, e.g., to compare mRNA expression levels of genes from organisms exposed to different conditions, researchers may report **fold-change**.

An example of calculation of fold change is rates of the expression from cells exposed to heavy metal divided by expression under basal conditions. Gene expression under different treatments may be evaluated by calculating fold-change as the log base 2 of the ratio of expression of a gene for one treatment divided by expression of the same gene from control conditions. Copper is an essential trace element, but excess exposure to copper is known to damage human health, including chronic obstructive pulmonary disease. One proposed mechanism is that cell injury promotes an epithelial-to-mesenchyme shift. In a pilot study we investigated gene expression changes by quantitative real-time polymerase chain reaction (qPCR) in a rat lung Type II alveolar cell line exposed to copper sulfate compared to unexposed cells. We recorded cycle threshold values, C_T , for each gene, where C_T is the number of cycles required for the fluorescent signal to exceed background levels; C_T is inversely proportional to amount of cDNA (mRNA) in the sample. Genes investigated were ECAD, FOXC2, NCAD, SMAD, SNAI1, TWIST, and VIM, with ATCB as reference gene. ECAD expression is marker of epithelial cells, whereas FOXC2, NCAD, SNAI1, TWIST, and VIM expression marker of mesenchymal cells. After calculating $2^{-\Delta\Delta C_T}$ values, **geometric means** of normalized values of three replicates each are shown in Table 6.2.1.

Note:

Logarithm transform is used because gene expression levels vary widely on the original scale and any log-transform will reduce the variability. log-base 2 is used for fold-change in particular because it is easy to interpret and provides symmetry (all log-transforms provide this symmetry). For example, $\log(1/2, 2)$ returns -1 , while $\log(2/1, 2)$ returns $+1$.

Thus, when using base 2, we see a decrease by half or doubling of original scale is a fold change of ± 1 . In contrast, $\log(1/2, 10)$ returns -0.301 , while $\log(2/1, 10)$ returns $+0.301$.

Table 6.2.1. Mean $2^{-\Delta\Delta C_T}$ and fold change of gene expression values from qPCR for several genes from a rat lung cell line.

	Control		Copper-sulfate	Fold change
ECAD	34.6	35.7	0.6	
NCAD	28.5	24.0	27.2	
SMAD	29.5	25.0	28.2	
SNAI1	25.5	28.1	0.2	
FOXC2	27.6	27.0	1.9	
VIM	23.1	16.4	134.4	
TWIST	25.1	22.9	5.6	

At face value, there appears to be some evidence that following a four-hour exposure to copper sulfate in media, the epithelial cell line adopted gene expression profile of mesenchyme-like cells. However, the weakness of fold-change is clear from Table 6.2.1: the quantity is sensitive to small values. ECAD expression in the cell line is low, thus the treated cells go through high numbers of PCR cycles (mean = 36) and control cells not much fewer (mean = 34.4).

Note:

Calculation of $2^{-\Delta\Delta C_T}$ is included. Geomean C_T were

	Control	CuSO ₄
ACTB	32.2	32.5
NCAD	28.5	24.0

For control cells, $\Delta C_{T\text{ control}} = C_{T\text{ GO1}} - C_{T\text{ Ref}} = 28.5 - 32.2 = -3.7$

For treatment cells, $\Delta C_{T\text{ treatment}} = C_{T\text{ GO1}} - C_{T\text{ Ref}} = 24.0 - 32.5 = -8.5$

and $\Delta\Delta C_T = \Delta C_{T\text{ treatment}} - \Delta C_{T\text{ control}} = -4.8$

$2^{-\Delta\Delta C_T} = 28$

Table value differs by rounding

Rates

Rates are a class of ratios in which the denominator is some measure of time. For example, the four year graduation rate of some Hawaii universities are shown in Table 6.2.2.

Table 6.2.2. Percent students graduation with bachelor's degrees within four years or six years (cohort 2014, data source [NCES.ed.gov](https://nces.ed.gov)).

School	Private/Public	Four-year, Percent graduation	Six-year, Percent graduation
Chaminade University	Private (non-profit)	43	58
Hawaii Pacific University	Private (non-profit)	31	46
University of Hawaii – Hilo	Public	15	38
University of Hawaii – Manoa	Public	35	62

School	Private/Public	Four-year, Percent graduation	Six-year, Percent graduation
University of Hawaii – West Oahu	Public	16	39
University of Phoenix	Private (for profit)	0	19

Examples of rates

Rates are common in biology. To name just a few:

- Basal metabolic rate (BMR), often measured by indirect calorimetry, reported in units kilo Joules per hour.
- Birth and death rates, components of population growth rate.
- Phred quality score, error rates of incorrectly called nucleotide bases by the sequencer
- Growth rate, which may refer to growth of the individual (somatic growth rate), or increase of number of individuals in a population per unit time
- Molecular clock hypothesis, rate of amino acid (protein) or nucleotide (DNA) substitution is approximately constant per year over evolutionary time.

Proportions

Proportions are also ratios, but they are used to describe one part to the whole. For example, 902 women (self-reported) day undergraduate students enrolled in 2014 at Chaminade University in Honolulu, Hawaii.

Practice: Given that the total enrollment for Chaminade in 2014 was of 1,326, calculate the proportion of female students to the total student body.

$$\frac{902}{1326} = 0.68$$

Comparing proportions

In some cases you may wish to compare two proportions or two ratios. The hypothesis tested is the difference between the two ratios, and the test is if the confidence interval of the difference includes zero. If it does, then we would conclude there is no statistical difference between the two proportions. In R, use the `prop.test` function. For example, 63 women were on team sport rosters at Chaminade in 2014, a proportion of 59% of all student athletes ($n = 106$). Recall from the example above that women were 68% of all students at Chaminade University. Title IX compliance requires that a university “maintain policies, practices and programs that do not discriminate against anyone on the basis of gender” (NCAA, <http://www.ncaa.org/about/resources/inclusion/title-ix-frequently-asked-questions>). In terms of athletic programs, then, universities are required to provide participation opportunities for women and men that are substantially proportionate to their respective rates of enrollment of full-time undergraduate students (NCAA, <http://www.ncaa.org/about/resources/inclusion/title-ix-frequently-asked-questions>)

Consider Chaminade University: Is there a statistical difference between proportion of women athletes and their proportion of total enrollment? We introduce statistical inference in [Chapter 8](#), but for now, this is a test of the null hypothesis that the difference between the two proportions is zero.

At the R prompt type (remember, anything after the # sign is a comment and ignored by R).

```
women = c(62,902) #where 62 is the number of women athletes and 902 is the number of
students = c(106,1326) #106 is the number of student athletes and 1326 is all studen
prop.test(women,students) #the default is a two-tailed test, i.e., no group differen
```

And R returns

```
2-sample test for equality of proportions with continuity correction
data: women out of students
```

```
X-squared = 3.6331, df = 1, p-value = 0.05664
alternative hypothesis: two.sided
95 percent confidence interval:
 -0.197532407 0.006861073
sample estimates:
 prop 1 prop 2
0.5849057 0.6802413
```

What is the conclusion of the test?

When you compare two groups, you're asking whether the two groups are equal (the null hypothesis). Mathematically, that's the same as saying the difference between the two groups is equal to zero.

First check the lower and upper limits of the confidence interval. A **confidence interval** is one way to report a range of plausible values for an estimate (see [Ch 7.6 – Confidence intervals](#)). It's called a confidence interval because a probability is assigned to the range of values; a 95% confidence interval is interpreted as we're 95% certain the true population value is somewhere between the reported limits. For our Chaminade University Title IX question, recall that we are asking whether the value of zero is included. The lower limit was -0.1975 and some change; the upper limit was 0.0068 and some change. Thus, zero is included and we would conclude that there was no statistical difference between the two proportions.

The second relevant output to look at is the **p-value**, or **probability value**. If the p-value is less than 5%, we typically reject the tested hypothesis. We will talk more about p-values and their relationship to inference testing in [Chapter 8](#); for now, pay attention to the confidence interval (introduced in [Chapter 3.4](#)); if zero is included, then we conclude no substantial differences between the two proportions.

Indexes

Indexes are composite statistics that combine indicators. Indexes are common in business and economics, e.g., Dow Jones Industrial average combines stock prices from 30 companies listed on the New York Stock Exchange.

Some indexes presented in this book include

- Grade point average
- Body Mass Index (BMI)
- Comet assay indexes (tail intensity, tail length, tail moment) are used to assess DNA damage among organisms exposed to environmental contaminants (e.g., Mincarelli et al., 2019).
- Encephalization index, ratio of brain to body weight among species. Used to compare cognitive abilities.

Scales

Agreement scales for surveys, e.g., **Likert scale** or sliding scale (Sullivan and Artino 2013). For example, after learning about [Theranos](#), students were asked:

How serious is this violation in your opinion (on a 5-point scale)?

Not serious 0	Slightly serious 0	Moderately serious 2	Serious 4	Very serious 19
------------------	-----------------------	-------------------------	--------------	--------------------

Although an intuitive measure, how fast an individual can run is challenging to determine because it is difficult to ensure that an individual's performance is at physiological maximum. Measures of performance capacity that involve behavior (motivation) can be particularly challenging, which may lead to the use of a **race quality** scale (eg., binary scale "good" or "bad" Husak et al 2006).

These examples reflect ordinal scales. Many of the nonparametric tests discussed in [Chapter 15](#) are suitable for analysis of scales.

Limitations of ratios

Although the indexes may be easy to communicate, statistically, indexes have many drawbacks. Chief among these is that variation in ratios may be due to change in numerator or denominator. Ratios and any index calculated by combining ratios seem simple enough, but have complicated statistical properties. Over the years, several authors have made critical suggestions for use of ratios

and indexes. Some key references are Packard and Boardman (1988), Jasienski and Oikos (1999), Nee et al (2005), and Karp et al (2012). For example, ratios, computing trait value by body weight, are often used to compare some trait among individuals or species that differ in body size. However, this **normalization** attempt only removes the covariation between size and the trait if there is a 1:1 relationship between size and the trait. More typically, relationship between the trait and body size is allometric, i.e., the slope is not equal to one. Thus, ratio will over-correct for large size and under-correct for small size. The proper solution is to conduct the comparison as part of an analysis of covariance (ANCOVA, see [Chapter 17.6](#)).

Example

Which is the safer mode of travel: car or airplane?

The following discussion covers travel safety in the United States of America for a typical year, 2000*.

Note:

*Note that the following discussion excludes the 241 airline passenger deaths associated to the terrorist attacks of September 11, 2001 in the USA; the NTSB also "...exclude(s illegal acts) for the purpose of accident rate computation." It also does not include considerations of 2020–2021 and effects of the COVID-19 pandemic on numbers of flights. The purpose of this discussion is not to convince you about the safety of modes of travel. Moreover, the following analysis is not necessarily the proper way to frame or analyze risk, but, rather, the purpose of this discussion is to highlight the impact of assumptions on estimating risk.

Between 2000 and 2023, there were 779 deaths associated with accidents of major air carriers in the USA. Year 2009 was the last multiple-casualty crash of a major U.S. carrier ([Colgan Air Flight 3407](#)); between 2010 and 2021, two fatal accidents, two fatalities were reported.

We've all heard the claim that it's much safer to fly with a major airline than it is to travel by car (e.g., [1 January 2012 article in online edition of San Francisco Chronicle](#)). There are a variety of arguments, but one statistical argument goes as follows. In 2000 in the United States, 638,902,993 persons traveled by major air carrier, whereas there were 190,625,023 licensed drivers. In 2000, 92 persons died in air travel (again, major carriers only), whereas 37,526 persons died in vehicle crashes (includes drivers and passengers). Thus, the risk of dying in air travel is given as the proportion $\frac{92}{638902993}$, or 1.44×10^{-7} (0.000014%), whereas the comparable proportion for death by motor vehicle is $\frac{37526}{190625023}$, or 1.97×10^{-4} (0.0197%).

In other words, we can expect one death (actually 1.4) for every ten million airline passengers, but 20 deaths (actually 19.7) for every one hundred thousand licensed drivers. Thus, flying is a thousand times safer than driving (actual result 1,367 times; divide the rate of motor vehicle-caused deaths for licensed drivers by the rate for airlines). Proportions are hard to compare sometimes, especially when the **per capita** numbers differ (ten million vs. 100,000 in this case).

We can put the numbers onto a **probability tree** and get a sense of what we are looking at.

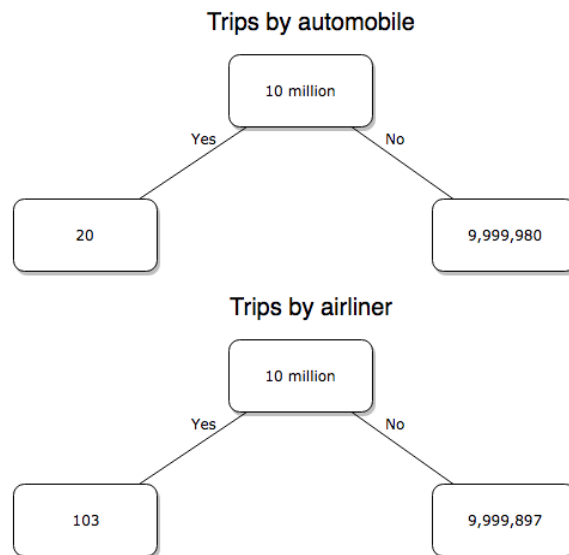


Figure 6.2.1: A **probability tree** to help visualize comparison of deaths (“yes”) by car travel and by airline travel in the United States for the year 2000.

Comparing rates and proportions

Without going into the details, we will do so in [Chapter 9: Inferences on Categorical Data](#), comparing two rates is a **chi-square**, χ^2 , **contingency table** type of problem. More specifically, however, it is a **binomial** problem ([Chapter 3.1](#), [Chapter 6.5](#)); there are two outcomes, death or no death, and we can describe how likely the event is to occur as a proportion. Because the numbers are large, we can use rely on the **normal distribution** for comparing the two proportions. We’ll explain this more in the next chapters, but for now it may be enough to present the equation for the comparison of two proportions under the assumption of normality, **proportion z test**.

$$z = \frac{(\hat{p}_1 - \hat{p}_2) - 0}{\sqrt{\hat{p}(1 - \hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

and the null hypothesis (see [Chapter 8](#)) tested as that the two proportions are equal. This may be written as

$$H_0 : p_1 - p_2 = 0$$

We can assign statistical significance to the differences in events for the two modes of travel under this set of assumptions.

`Rcmdr` has a nice menu-driven system for comparing proportions, but for now I will simply list the R commands.

At the R prompt, type each line then submit the command.

```
total = 100000000
prop.test(c(19700,14),c(total,total))
```

And the R output is:

```
prop.test(c(19700,14),c(total,total))

2-sample test for equality of proportions with continuity
correction

data:  c(19700, 14) out of c(total, total)
X-squared = 19658, df = 1, p-value < 2.2e-16
alternative hypothesis: two.sided
```

```
95 percent confidence interval:
 0.0001940984 0.0001996216
sample estimates:
  prop 1    prop 2
1.97e-04 1.40e-07
```

There's a bit to unpack here. R is consistent; when it reports results of a statistical test, it typically returns the value of the test statistic ($\chi^2 = 19658$), the degrees of freedom for the test ($df = 1$), and the p-value ($< 2.2e-16$).

 Note:

The confidence intervals reported by `prop.test()` were calculated by the **Wilson Score method**, not the **Wald method**. While both are parametric tests and therefore sensitive to departures from normality (see [Chapter 13.3](#)), formulation of Wilson score method makes fewer assumptions (involving approximations of the population proportions) and therefore is considered more accurate.

By convention in statistics, if a p-value, where “p” stands for probability, is less than 5%, we would say that our results are statistically significant from the null hypothesis. Looks pretty convincing to me; the difference of 19,700 deaths compared to 14 deaths is clearly different by any criterion and by the results of the statistical test, the p-value is several **orders of magnitude** smaller than 5%.

 Note:

Order of magnitude generally refers to differences in multiples of ten, logarithmic: a difference of one order of magnitude is the number multiplied by 10^1 , three orders of magnitude is the number multiplied by 10^3 , and so on.

Safer to fly. By far, not even close. And similar conclusions would be reached if we compare different years, or averages over many years, or if we used a different way to express the amount of travel (e.g., miles/year) by these modes of transportation.

Are you convinced, really? Is it safer to fly?

Let's try a little statistical reasoning — what **assumptions** did I make to do these calculations? We recognize immediately that many more people travel by car: that there are way more cars being driven than there are airline planes being flown. The question then is, have we properly adjusted for this difference? Here are a few considerations. My source for the numbers is the *NTS 2001* book published by the U.S. Department of Transportation (www.dot.gov). We are conducting a risk analysis, and the first step is to make sure that we are comparing “apples with apples.” Here are two alternative solutions that at least compare, “Red Delicious” apples with “Macintosh” apples.

Option 1

There are many, many more licensed drivers than there are licensed commercial airline pilots. The standard comparison offered in the background above compared deaths per licensed car driver, but a different metric for air travel, the rate per passenger. This isn't as bad of a comparison as it may seem — after all, the majority of deaths in car accidents are of the driver themselves. But it isn't that hard to make the direct comparison — just find out how many commercial pilots there are — a direct comparison with licensed car drivers (stated above as 190,625,023). From the FAA we see that in 2009 there were 125,738 persons with commercial certificates. Since there are only 20 major airline carriers in the United States now (a few more were active in 2000, but we'll put this aside), the number of licenses is an overestimate of the actual number we want — how many pilots of commercial airlines — but let's use this number for starters. After all, just because a person has a drivers license doesn't mean they drive or ride in a car.

Number of deaths/yr: Let's use 2000 data, a typical year prior to 9/11 (and excluding the Covid-19 pandemic). Airlines: 92 deaths; motor vehicles (includes passenger cars, trucks, etc., but not motorcycles): 37,526 deaths (drivers = 25,567; passengers = 10,695; 86 others).

Which mode of travel is riskier? I get a rate a rate of 7.3×10^{-4} deaths per commercial pilot, compared to a rate for car drivers of 1.97×10^{-4} deaths.

To summarize what we have so far, I get a result that suggests car travel is almost four times safer:

$$\frac{7.3 \times 10^{-4}}{1.97 \times 10^{-4}} \rightarrow \frac{7.3}{1.97} = 3.7$$

then traveling by commercial airliner. In whole numbers, these results translate to seven deaths for every 10,000 commercial pilots compared to two deaths for every 10,000 licensed car drivers.

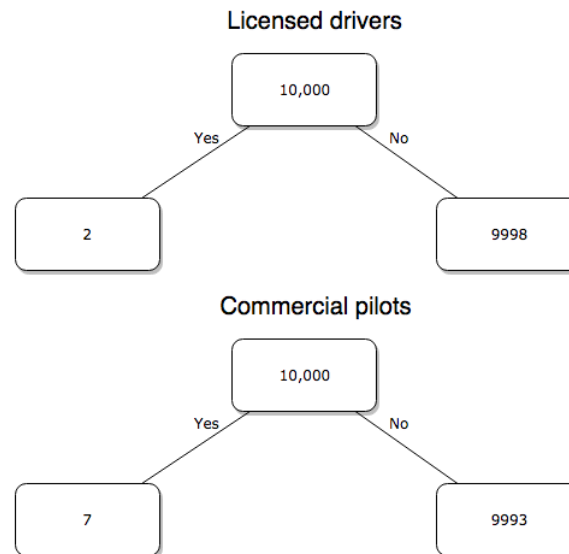


Figure 6.2.2: Comparing totals of deaths adjusted by numbers of licensed drivers and by licensed commercial airline pilots in the United States.

R work follows. Enter and submit each command on a separate line in the script window

```
total = 10000
prop.test(c(2,7),c(total,total))
```

And the R output

```
prop.test(c(2,7),c(total,total))

2-sample test for equality of proportions with continuity
correction

data:  c(2, 7) out of c(total, total)
X-squared = 1.7786, df = 1, p-value = 0.1823
alternative hypothesis: two.sided
95 percent confidence interval:
 -0.001187816  0.000187816
sample estimates:
prop 1 prop 2
 2e-04  7e-04
```

What's happened? The p-value (0.1823) is not less than 5%, and so we would conclude under this scenario that there is no difference between the proportions of deaths between the two modes of travel. Let's keep going.

Option 2

There are many, many more cars on the road then there are airplanes flying commercial passengers. The standard comparison offered in the background information above identified death rates per individual driver, but used a different metric for airline

travelers (number of deaths per passenger), which confuses individuals with travelers: what we need is the number of individuals that traveled by airliner, not the total number of passengers (which is many times higher, because of repeat flyers). How can we make a fair comparison for the two modes of travel? Most people never fly, whereas most people drive (or ride in a car) frequently in the United States. To me, risk of travel might be better expressed in terms of a per trip rate. I want to know, what are my chances of dying each time I get into my car versus each time I fly on a commercial jet in the United States?

Number of trips/yr. For airlines, I use the number of departures (in 2000 this was 8,951,773). But for cars, we need to decide how to get a similar number. It's not available directly from the DOT (and would be difficult to get — studies with randomly selected drivers can yield as many as 5 trips per day for licensed drivers). I took the number of licensed drivers and bound the problem — at the low end, let's say that only 2 trips per week (e.g., 50 weeks) are taken by licensed drivers (100 trips); at the upper end, let's take 2 trips per day per week, or 500 trips/year. Thus, at the low end, we have 1.91×10^{10} trips per year; at the upper end, 9.53×10^{10} trips per year.

Which mode of travel is riskier? Using the number of deaths/yr listed above in Option 1, I get a rate of 1.03×10^{-5} deaths per trip for air carriers compared to a rate of 1.97×10^{-6} deaths per trip for cars (lower bound) or 3.9×10^{-7} deaths per trip for cars (upper bound). Here's what the numbers look for in a tree (taking the lower number of trips per year for cars).

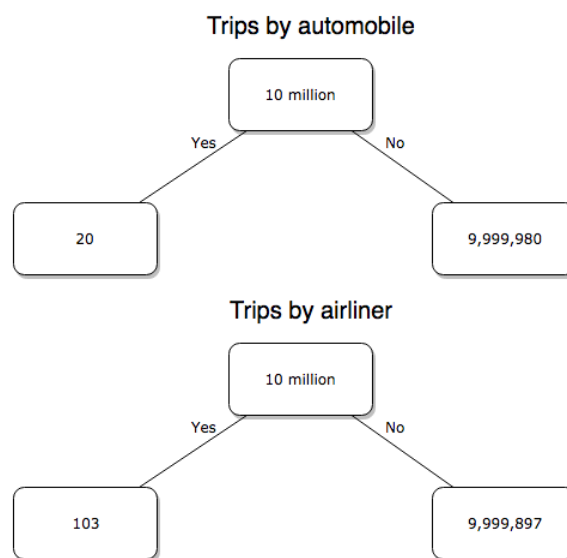


Figure 6.2.3: Comparing totals of deaths adjusted by numbers of car trips and by numbers of airline trips in the United States.

R work follows:

```
total = 10000000
prop.test(c(20,103),c(total,total))
```

And the R output:

```
prop.test(c(20,103),c(total,total))

2-sample test for equality of proportions with continuity
correction

data:  c(20, 103) out of c(total, total)
X-squared = 54.667, df = 1, p-value = 1.428e-13
alternative hypothesis: two.sided
95 percent confidence interval:
 -1.057370e-05 -6.026305e-06
```

```
sample estimates:
  prop 1    prop 2 
2.00e-06 1.03e-05
```

Now we have another really small p-value (1.428×10^{-13}), which suggests a statistically significant difference between the modes of travel, but the difference in deaths is switched. I now have a result that suggests car travel is *much* safer than traveling with a commercial airliner! These calculations suggest that you are as much as 26 (upper bounds, five times for lower bounds) times more likely to die from a plane crash than you are behind the wheel. In whole numbers, these results indicate one death for every 100,000 airline flights compared to 1 death for every 500,000 (lower estimate) or 2,500,000 car trips!

Do I have it right and the standard answer is wrong? As Lee Corso says often on ESPN's College GameDay program, "Not so fast, my friend!" ([Wikipedia](#)). Mark Twain was right to hold the skeptic's view. Begin by listing the assumptions and by checking the logic of the comparisons (there are still holes in my logic!!). For one, if I am considering my risk of dying by mode of travel, it is far more likely that I will be in a car accident than I will an airline accident, simply because I don't travel by airline that much. When we consider **lifetime risk**, we can see why the assertion that it is "safer to fly than drive" is true — we're far more likely to belong to one of the reference populations involving automobiles (e.g., those who drive frequently, for many years) than we are to be among the frequent flyers reference populations.

Questions

1. Review and provide your own examples for

- index
- rate
- ratio
- proportion

2. Return to my story about travel safety, airlines vs cars: am I using "statistic" or "statistics?"

3. Like travel safety, we are often confronted by risk comparisons like the following: Which animal is more deadly to humans, dogs or sharks? Between the two, which lead to more hospitalizations in the United States? Work through your assumptions and use results from the [International Shark Attack file](#).

- If a person lives in Nebraska, and never visits the ocean, how does a "shark attack" risk analysis apply? Is it a fair comparison to make between dog attacks and shark attacks? Why or why not.

4. Go to cappex.com/colleges and update institutional (gift) aid offered by Chaminade and HPU. Compare to University of Hawaii-Manoa.

This page titled [6.2: Ratios and probabilities](#) is shared under a [CC BY-NC-SA 4.0](#) license and was authored, remixed, and/or curated by [Michael R Dohm](#) via [source content](#) that was edited to the style and standards of the LibreTexts platform.