Hypothesis Testing

Measuring the consistency between a model and data

11/04/2022

Classical statistics features two dominant methods for using a sample of data to make an inference about a more general process. The first is the confidence interval.

The goal of a confidence interval is to express the uncertainty in an estimate of a population parameter

The second classical method of generalization is the hypothesis test. The hypothesis test takes a more active approach to reasoning: it posits a specific explanation for how the data could be generated, then evaluates whether or not the observed data is consistent with that model. The hypothesis test is one of the most common statistical tools in the social and natural sciences, but the reasoning involved can be counter-intuitive. Pay attention.

Case study: The United States vs Kristen Gilbert

In 1989, fresh out of nursing school, Kristen Gilbert got a job at the VA Medical Center in Northampton, Massachusetts, not far from where she grew up[1]. Within a few years, she became admired for her skill and competence.

Gilbert's skill was on display whenever a "code blue" alarm was sounded. This alarm indicates that a patient has gone into

[1] This case study appears in *Statistics in the Courtroom: United States v. Kristen Gilbert* by Cobb and Gelbach, published in *Statistics: A Guide to the Unknown* by Peck et. al.

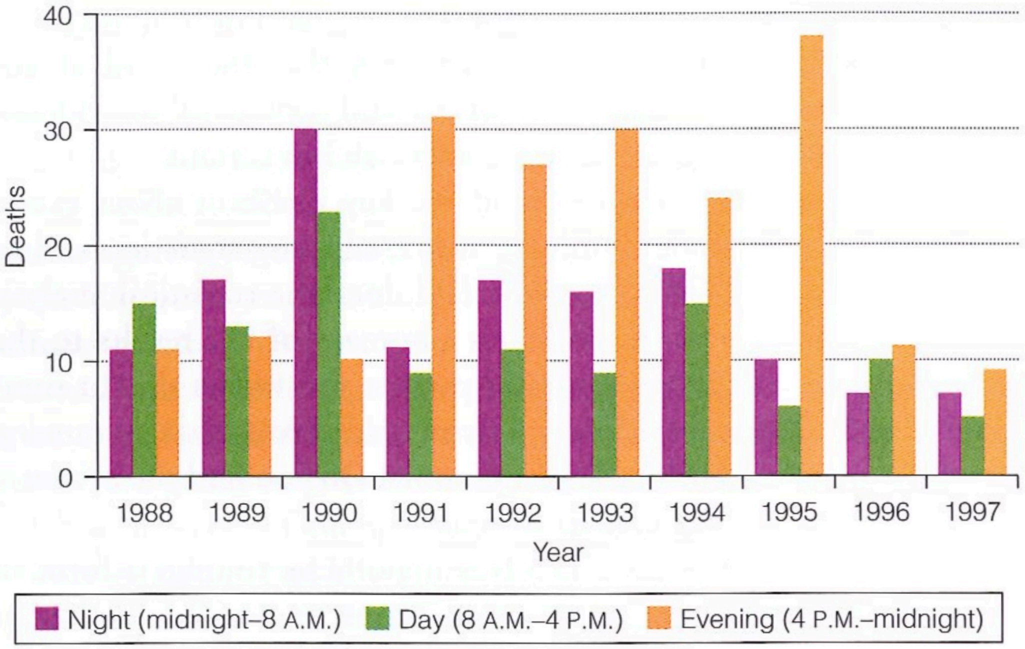
cardiac arrest and must be addressed quickly by administering a shot of epinephrine to restart the heart. Gilbert developed for a reputation for her steady hand in these crises.

By the mid-1990s, however, the other nurses started to grow suspicious. There seemed to be a few too many code blues during Gilbert's shifts. The staff brought their concerns to the VA administration, who brought in a statistician to evaluate the data.

The Data

The data that the VA provided to the statistician contained the number of deaths at the medical center over the previous 10 years, broken out by the three shifts of the days: night, daytime, and evening. As part of the process of exploratory data analysis, the statistician constructed a plot.

This visualization reveals several striking trends. Between 1990 and 1995, there were dramatically more deaths than the years before and after that interval. Within that time span, it was the evening shift that had most of the deaths. The exception is 1990, when the night and daytime shifts had the most deaths.



So when was Gilbert working? She began working in this part of the hospital in March 1990 and stopped working in February 1996. Her shifts throughout that time span? The evening shifts. The one exception was 1990, when she was assigned to work the night shift.

This evidence is compelling in establishing an association between Gilbert and the increase in deaths. When the district attorney brought a case against Gilbert in court, this was the first line of evidence they provided. In a trial, however, there is a high burden of proof.

Could there be an alternative explanation for the trend found in this data?

The role of random chance

Suppose for a moment that the occurrence of deaths at the hospital had nothing to do with Gilbert being on shift. In that case we would expect that the proportion of shifts with a death would be fairly similar when comparing shifts where Gilbert was working and shifts where she was not. But we wouldn't expect those proportions to be *exactly* equal. It's reasonable to think that a slightly higher proportion of Gilbert's shifts could have had a death just due to random chance, not due to anything malicious on her part.

So just how different were these proportions in the data? The plot above shows data from 1,641 individual shifts, on which three different variables were recorded: the shift number, whether or not there was a death on the shift, and whether or not Gilbert was working that shift.

```
# A tibble: 1,641 x 3
   shift death staff
   <dbl> <chr> <chr>
 1   626 no    no_gilbert
 2   590 no    no_gilbert
 3  1209 no    no_gilbert
 4  1122 no    no_gilbert
 5   622 no    no_gilbert
 6  1536 no    no_gilbert
 7  1472 no    no_gilbert
 8   214 no    gilbert
 9   277 yes   no_gilbert
10  1332 no    no_gilbert
# ... with 1,631 more rows
```

Using this data frame, we can calculate the sample proportion of shifts where Gilbert was working (257) that had a death (40) and compare them to the sample proportion of shifts where Gilbert was *not* working (1384) that had a death (34).

$$\widehat{p}_{gilbert} - \widehat{p}_{no\_gilbert} = \frac{40}{257} - \frac{34}{1384} = .155 - .024 = .131$$

A difference of .131 seems dramatic, but is that within the bounds of what we might expect just due to chance? One way to address this question is to phrase it as: if in fact the probability of a death on a given shift is independent of whether

A note on notation: it's common to use $\widehat{p}$ ("p hat") to indicate that a proportion has been computed from a sample of data.

or not Gilbert is on the shift, what values would we expect for the difference in observed proportions?

We can answer this question by using simulation. To a simulate a world in which deaths are independent of Gilbert, we can

1. Shuffle (or permute) the `death` variable in the data frame to break the link between that variable and the `staff` variable.
2. Calculate the resulting difference in proportion of deaths in each group.

The rationale for shuffling the columns is that if in fact those two columns are independent of one another, then it was just random chance that led to a value of one variable landing in the same row as the value of the other variable. It could just as well have been a different pairing. Shuffling captures another example of the arbitrary pairings that we could have observed if the two variables were independent of one another[2].

By repeating steps 1 and 2 many many times, we can build up the full distribution of the values that this difference in proportions could take.

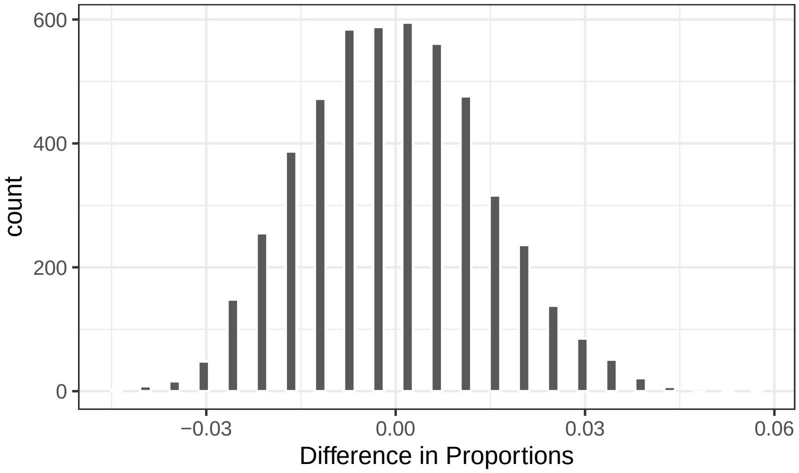
```r
library(infer)
set.seed(40215)
read_csv("https://www.dropbox.com/s/yj3grtilupyj9pv/code_blue.csv?dl=1")
```

```
# A tibble: 1,641 x 3
   shift death staff
   <dbl> <chr> <chr>
 1     1 yes   gilbert
 2     2 yes   gilbert
 3     3 yes   gilbert
 4     4 yes   gilbert
 5     5 yes   gilbert
 6     6 yes   gilbert
 7     7 yes   gilbert
```

[2] The technical notion that motivates the uses of shuffling is a slightly more general notion than independence called exchangability. The distinction between these two related concepts is a topic in a course in probability.

```
 8     8 yes   gilbert
 9     9 yes   gilbert
10    10 yes   gilbert
# ... with 1,631 more rows
```

```
null <- code_blue %>%
  specify(response = death, explanatory = staff, success = "yes") %>%
  hypothesize(null = "independence") %>%
  generate(reps = 5000, type = "permute") %>%
  calculate(stat = "diff in props")
null %>%
  ggplot(aes(x = stat)) +
  geom_bar(col = "white", bins = 23) +
  theme_bw() +
  labs(x = "Difference in Proportions")
```

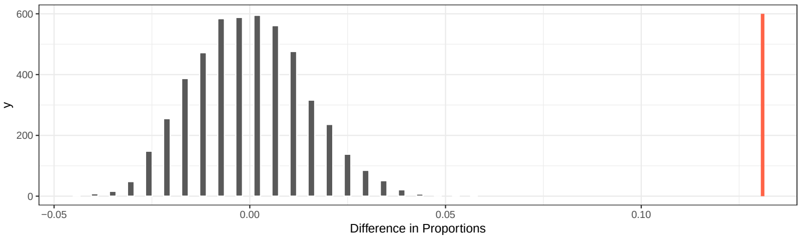


As expected, in a world where these two variables are independent of one another, we would expect a difference in proportions around zero. Sometimes, however, that statistic might reach values of +/- .01 or .02 or rarely .03. In the 500 simulated statistics shown above, however, none of them reached beyond +/- .06.

So if that’s the range of statistics we would expect in a world where random chance is the only mechanism driving the difference in proportions, how does it compare to the world that we actually observed? The statistic that we observed in the data was .131, more than *twice* the value of the most extreme statistic observed above.

To put that into perspective, we can plot the observed statistic as a vertical line on the same plot.

```
null %>%
  ggplot(aes(x = stat)) +
  geom_bar(col = "white", bins = 23) +
  theme_bw() +
  labs(x = "Difference in Proportions") +
  annotate("segment", x = .131, xend = .131,
           y = 0, yend = 600, color = "tomato", lwd = 1.5)
```



The method used above shows that the chance of observing a difference of .131 is incredibly unlikely if in fact deaths were independent of Gilbert being on shift. On this point, the statisticians on the case agreed that they could rule out random chance as an explanation for this difference. Something else must have been happening.

Elements of a Hypothesis Test

The logic used by the statisticians in the Gilbert case is an example of a hypothesis test. There are a few key components common to every hypothesis test, so we'll lay them out one-by-one.

A hypothesis test begins with the assertion of a null hypothesis.

Null Hypothesis A description of the chance process for generating data.

It is common for the null hypothesis to be that nothing interesting is happening or that it's business as usual, a hypothesis that statisticians try to refute with data. In Gilbert case, this could be described as "The occurrence of a death is independence of the presence of Gilbert" or "The probability of death is the same whether or not Gilbert is on shift" or "The difference in the probability of death is zero, when comparing shifts where Gilbert is present to shifts where Gilbert is not present". Importantly, the null model describes a possible state of the world, therefore the latter two versions are framed in terms of *parameters* (p for proportions) instead of observed *statistics* ($\hat{p}$).

The hypothesis that something indeed is going on is usually framed as the alternative hypothesis.

Alternative Hypothesis An alternative to the null hypothesis. Often: "the difference is real".

In the Gilbert case, the corresponding alternative hypothesis is that there is "The occurrence of a death is *dependent* on the presence of Gilbert" or "The probability of death is *different* whether or not Gilbert is on shift" or "The difference in the probability of death is *non-zero*, when comparing shifts where Gilbert is present to shifts where Gilbert is not present"

In order to determine whether the observed data is consistent with the null hypothesis, it is necessary to compress the data down into a single statistic.

Test Statistic A numerical summary of the observed data that bears on the null hypothesis. Under the null hypothesis it has a sampling distribution (also called a "Null Distribution").

In Gilbert's case, a difference in two proportions, $\hat{p}_1 - \hat{p}_2$ is a natural test statistic and the observed test statistic was .131.

It's not enough, though, to just compute the observed statistic. We need to know how likely this statistic would be in a world where the null hypothesis is true. This probability is captured in the notion of a p-value.

p-value The chance of a test statistic as rare or even more rare than the one observed under the assumptions of the null hypothesis.

If the p-value is high, then the data is consistent with the null hypothesis. If the p-value is very low, however, there the statistic that was observed would be very unlikely in a world where the null hypothesis was true. As a consequence, the null hypothesis can be rejected as reasonable model for the data.

The p-value can be estimated using the proportion of statistics from the simulated null distribution that are as or more extreme than the observed statistic. In the simulation for the Gilbert case, there were 0 statistics greater than .131, so the estimated p-value is zero.

What a p-value is not

The p-value has been called the most used as well as the most abused tool in statistics.

1. The p-value is not the probability that the null hypothesis is true.

   This is one of the most common

Another check on your understanding of a p-value. A p-value is a probability, therefore it must between a number between 0 and 1. If you ever find yourself computing a p-value of -6 or 3.2, be sure to pause and revisit your calculations!

Summary