Hypothesis Testing

David Gerard

2017-11-02

Learning Objectives

- Hypothesis tests.
- Connection to confidence intervals.
- Section 4.3 of DBC

Motivation

- Often, scientists want to test for binary decisions.
- E.g. Does this gene impact height (Yes/No)
- E.g. Does broccoli cause cancer (Yes/No)
- E.g. Is Trump's phone source associated with negative words (Yes/No)?
- Today, we'll talk about making binary decisions in the context of a question on Old Faithful's reliability.

Hypothesis test

- A hypothesis test is an assessment of the evidence provided by the data in favor of (or against) some claim about the population.
- For example, suppose we perform a randomized experiment or take a random sample and calculate some sample statistic, say the sample mean.
- We want to decide if the observed value of the sample statistic is consistent with some hypothesized value of the corresponding population parameter.
- If the observed and hypothesized value differ (as they almost certainly will), is the difference due to an incorrect hypothesis or merely due to chance variation?

Old Faithful

- Old Faithful is a geyser in Yellowstone National Park that is known for erupting approximately once every hour.
- That is, the lore is that the average eruption time for Old Faithful is 60 minutes.
- We want to see if data corroborate this lore.

Old Faithful Dataset

Waiting time between eruptions and the duration of the eruption for the Old Faithful geyser in Yellowstone National Park, Wyoming, USA.

Data consist of two variables

- `duration` Eruption time in mins.
- `waiting` Waiting time to this eruption (in mins).

Old Faithful Dataset

```
library(tidyverse) ## for glimpse() function
library(MASS) ## contains geyser dataset
data("geyser")
glimpse(geyser)

Observations: 299
Variables: 2
$ waiting  <dbl> 80, 71, 57, 80, 75, 77, 60, 86, 77, 56...
$ duration <dbl> 4.017, 2.150, 4.000, 4.000, 4.000, 2.0...

waiting <- geyser$waiting
```

Hypotheses

Using these data, we wish to decide between one of two hypotheses:

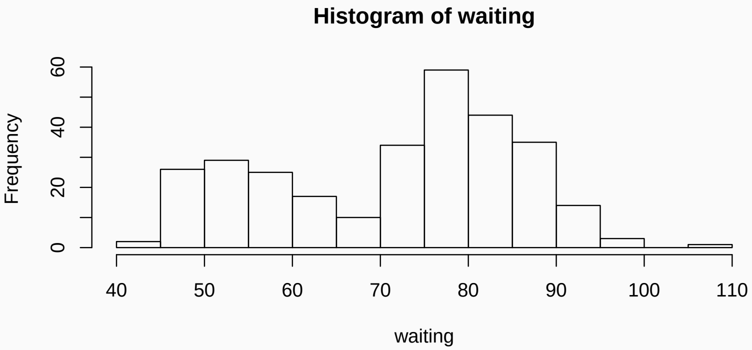
- H_0 The mean eruption time μ for Old Faithful is 60 minutes.
- H_A The mean eruption time μ for Old Faithful is not 60 minutes.
- Formulating different hypotheses is the first step in any testing scenario.

General Form of Hypotheses

- The null hypothesis H_0 is the statement being tested. Usually it states that the difference between the observed value and the hypothesized value is only due to chance variation. For example, $\mu = 60$.
- The alternative hypothesis H_A is the statement we will favor if we find evidence that the null hypothesis is false. It usually states that there is a real difference between the observed and hypothesized values. For example, $\mu \neq 60$, $\mu > 60$, or $\mu < 60$.
- A test is called
    - two-sided if H_A is of the form $\mu \neq 60$.
    - one-sided if H_A is of the form $\mu < 60$ or $\mu > 60$.

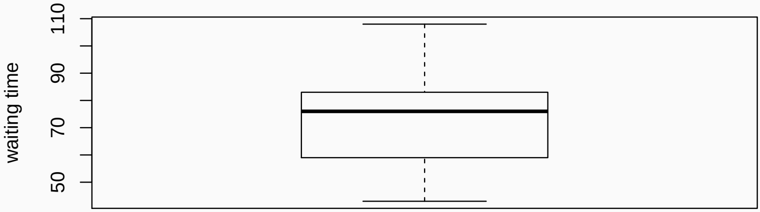
Some EDA i

```r
hist(waiting)
```



Some EDA ii

```
boxplot(waiting, ylab = "waiting time")
```



Some EDA iii

```r
summary(waiting)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
   43.0    59.0    76.0    72.3    83.0   108.0 
```

Conclusion from EDA

- EDA suggests $\mu \neq 60$, but again, this might be due to random variation.
- We need some formal way to evaluate the unlikeliness of the data we observe under H_0.

Confidence Intervals

Recall for large enough n

$$\bar{x} \sim N(\mu, \sigma^2/n).$$

- We used this result in the previous lecture to come up with 95% confidence intervals.
- That is, $\bar{x}$ is will only deviate from μ by more than 2 standard deviations in approximately 5% of repeated samples.
- So μ is between $\bar{x} - 2\sigma/\sqrt{n}$ and $\bar{x} + 2\sigma/\sqrt{n}$ in about 95% of repeated samples.
- So if our hypothesized μ (60 minutes) is outside of this interval, it is unlikely that $\mu = 60$.

Calculating CI

- $\bar{x} = 72.31$.
- $s = 13.89$.
- CI: $(70.74, 73.89)$.
- Since $60 \notin (70.74, 73.89)$, we are left with one of two conclusions:
    1. H_0 is true (so $\mu = 60$) and what we observed is an extremely rare event.
    2. H_A is true and $\mu \neq 60$.
- Since the data are unlikely to have been observed if H_0 were true, we reject H_0 and conclude that $\mu \neq 60$.

Type I Error

- What if H_0 were actually true?
- Recall that a 95% CI only covers the true μ in 95% of repeated samples.
- So the sample we actually observed could be one of the 5% of samples that misses the true μ and we incorrectly rejected H_0.
- When we incorrectly reject H_0 this is called (rather stupidly) a Type I error.
- Some call this, more intuitively, a false discovery or a false positive.
- If we used, instead of a 95% CI, a (1 - α)% CI, in what proportion of repeated samples would we make a Type I error?

Type II Error

- We could also have failed to reject H_0 when in fact H_0 is false.
- This is called a Type II Error, or a false negative.

Subtle Language

- We say "reject H_0" when we have evidence against H_0.
- We say "fail to reject H_0" when we do not have enough evidence against H_0.
- We generally never say "accept H_0".
- There are philosophical reasons for this: lack of evidence against a hypothesis is not the same as evidence for a hypothesis — e.g. a "not guilty" verdict in court does not mean "innocent".
- There are practical reasons for this: If a scientist wanted to publish a result, he could make his desired hypothesis H_0 and then collect a very small sample size. He would usually fail to reject H_0 and could publish a lot of bad papers.

More Formal Testing

Motivation

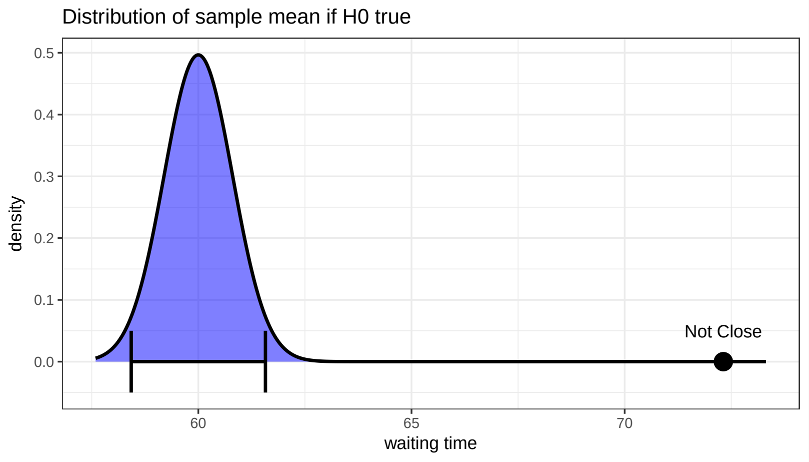
The CI approach to hypothesis testing is too course.

- If a hypothesized μ is just inside a 95% confidence interval, we want to say that we fail to reject H_0, but it was a close call.

Motivation

The CI approach to hypothesis testing is too course.

- If a μ so so far away from the boundary of the 95% CI, we want to say that H_0 is super super unlikely to be true.



p-value

p-value

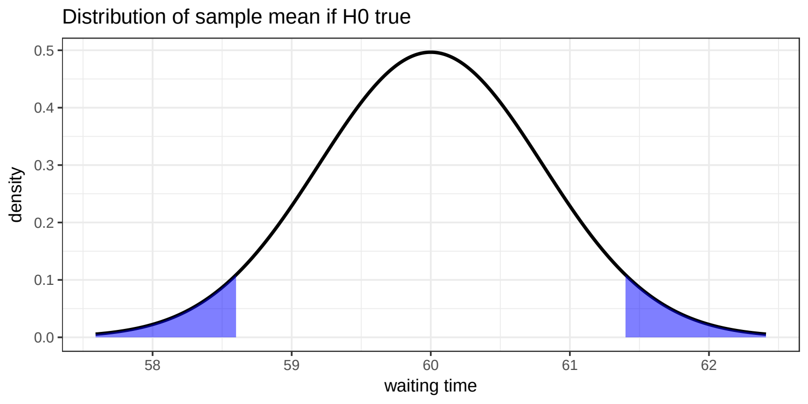
The p-value is the probability of observing data at least as favorable to the alternative hypothesis as our current data set, *if the null hypothesis were true*.

- A small p-value (close to 0) means that the data would be very unlikely under H_0, providing evidence for H_A.
- A large p-value (not close to 0) means that the data would be likely under H_0, *not* providing evidence for H_A.
- Generally, we reject H_0 if the p-value is below some level α. In this case, α is called the significance level of a test.

How do we caluclate a p-value?

We know the distribution of $\bar{x}$ under H_0, so we can calculate the probability of seeing data as extreme or more extreme than $\bar{x}$ under H_0 using normal probabilities.

E.g. If $\bar{x} = 61.4$, we would calculate these probabilities.



Why both tails?

- Recall that $H_A : \mu \neq 60$.
- The definition of a p-value is the probability of seeing something *as extreme or more extreme* (under the null) than what we saw.
- Since $\mu_0 - (\bar{x} - \mu_0)$ is as extreme as $\bar{x}$, we have to include this in our p-value calculation.

One sided hypothesis

If $H_A : \mu > 60$ and $\bar{x} = 61.4$.

One sided hypothesis

If $H_A : \mu < 60$ and $\bar{x} = 61.4$.

But we are in this case

How do we calculate these probabilities?

How do we calculate these probablities

- We have, under the null $\bar{X} \sim N(\mu, \sigma^2/n)$
- We want $Pr(\bar{X} > 61.4$ or $\bar{X} < 58.6)$ (since 58.6 is 1.4 away from 60, as is our (pretend) observed statistics 61.4).
- This is equal to $2Pr(\bar{X} > 61.4)$.
- We will insert $s = 13.8903$ for σ here.

```
2 * pnorm(q = 61.4, mean = 60, sd = 13.89 / sqrt(299),
          lower.tail = FALSE)
```

```
[1] 0.08136
```

Another way

We could also use this fact

Standard Normal

Let $X \sim N(\mu, \sigma^2)$. Let $Z = \frac{X-\mu}{\sigma}$. Then $Z \sim N(0, 1)$. The normal distribution with mean 0 and standard deviation 1 is sometimes called the standard normal distribution.

Using Standard Normal

In which case,

$$\begin{aligned} Pr(|\bar{X} - 60| > 1.4) &= Pr\left(\left|\frac{\bar{X} - 60}{13.89/\sqrt{299}}\right| > 1.743\right) \\ &= Pr(|Z| > 1.743), \end{aligned}$$

where $Z \sim N(0, 1)$.

```
2 * pnorm(-1.743)
```

```
[1] 0.08133
```

Conclusion

- 61.4 is a made-up value. But if it were real, we might choose a significance level of $\alpha = 0.05$.
- In which case, since $0.0813 > 0.05$, we would *fail to reject H_0* and say that we do not have enough evidence to conclude that Old Faithful erupts differently than once every hour.
- Why $\alpha = 0.05$? **NO REASON**. But everyone in the entire world uses $\alpha = 0.05$.

Real data

The value 61.4 was made up. Let's calculate the p-value given our real observation of 72.31.

```
xbar <- mean(waiting)
s    <- sd(waiting)
n    <- length(waiting)
z    <- (xbar - 60) / (s / sqrt(n))
2 * pnorm(-abs(z))

[1] 4.837e-53
```

Conclusion

- Since $4.8365 \times 10^{-53} << 0.05$, we strongly reject H_0 and conclude that Old Faithful does not on average erupt once an hour.

How to interpret the significance level

- Suppose P is the p-value we obtain. Then P is itself a random variable that has a distribution.
- Given a significance level, α, then one can show that, under the H_0, $Pr(P \leq \alpha) = \alpha$.
- That is, if we reject H_0 whenever $P < \alpha$, then we would expect a Type I error rate of α under the null.
- A larger significance level α means that we have a larger Type I error rate, but a smaller Type II error rate.
- A smaller α means that we have a smaller Type I error rate but a larger Type II error rate.
- We generally only control for Type I error rate (by setting α).

Summary of p-value for means and further thoughts

Step 1

Formulate the null hypothesis and the alternative hypothesis

- The **null hypothesis** H_0 is the statement being tested. Usually it states that the difference between the observed value and the hypothesized value is only due to chance variation. For example, $\mu = 60$ minutes.
- The **alternative hypothesis** H_a is the statement we will favor if we find evidence that the null hypothesis is false. It usually states that there is a real difference between the observed and hypothesized values.
  For example, $\mu \neq 60$, $\mu > 60$, or $\mu < 60$.

A test is called

- **two-sided** if H_A is of the form $\mu \neq 60$.
- **one-sided** if H_A is of the form $\mu > 60$, or $\mu < 60$.

Step 2

Calculate the **test statistic** on which the test will be based.

The test statistic measures the difference between the observed data and what would be expected *if* the null hypothesis were true.

Our goal is to answer the question, "How many standard errors is the observed sample value from the hypothesized value (under the null hypothesis)?"

For the Old Faithful example, the test statistic is

$$z = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} = \frac{72.31 - 60}{13.89/\sqrt{299}} = 15.3298$$

Step 3

Find the *p*-**value** of the observed result

- The *p*-value is the probability of observing a test statistic *as extreme or more extreme than actually observed*, assuming the null hypothesis H_0 is true.
- The smaller the *p*-value, the stronger the evidence *against* the null hypothesis.
- if the *p*-value is as small or smaller than some number α (e.g. 0.01, 0.05), we say that the result is **statistically significant** at level α.
- α is called the **significance level** of the test.

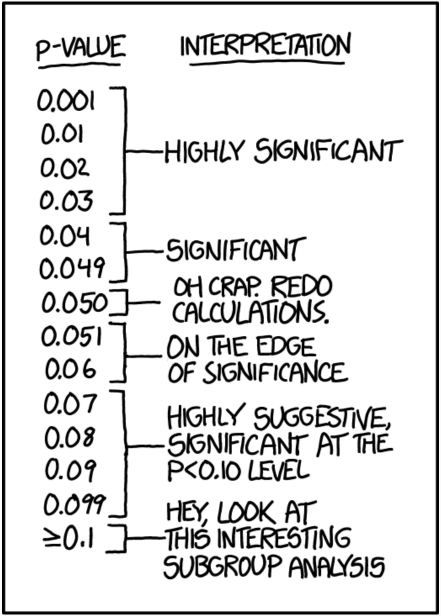
In the case of the Old Faithful example, $p = 4.8365 \times 10^{-53}$ for a two-sided test.

How to calculate p-values

For $Z \sim N(0, 1)$, the p-values for different alternative hypotheses:

- $H_A : \mu > \mu_0$ – p-value is $P(Z \geq z)$ (area of right-hand tail)
- $H_A : \mu < \mu_0$ – p-value is $P(Z \leq z)$ (area of left-hand tail)
- $H_A : \mu \neq \mu_0$ – p-value is $2P(Z \geq |z|)$ (area of both tails)

How to interpret *p*-values



Test interpretations

Saying that a result is statistically significant does not signify that it is large or necessarily important. That decision depends on the particulars of the problem. A statistically significant result only says that there is substantial evidence that H_0 is false. Failure to reject H_0 does not imply that H_0 is correct. It only implies that we have insufficient evidence to conclude that H_0 is incorrect.

Correct/Incorrect interpretation of Hypothesis tests?

1. The p-value is the probability of seeing data that supports the alternative hypothesis as strong or stronger than what we saw.
2. The p-value is the probability that the null hypothesis is correct. A smaller p-value means that the null is less probable and so we may reject it in favor of the alternative.
3. A large p-value is strong evidence in favor of the null hypothesis.
4. If we rejected H_0, then the null hypothesis is totally not true.
5. If $\alpha = 0.05$, then we would expect about 1 study in 20 to incorrectly reject the null hypothesis.

Proportions

Proportions

What if we have 0/1 (Bernoulli) data? E.g. the CLOUDS variable from the Bob Ross dataset.

$$Z_i = \begin{cases} 1 & \text{if a cloud is in the painting} \\ 0 & \text{if a cloud is not in the painting.} \end{cases}$$

Then the proportion of clouds is itself a mean

$$\hat{p} = \bar{x} = \frac{1}{n} \sum_i z_i = 0.4442$$

Using CLT

Say we wanted to test the hypothesis that Bob uses clouds less than 50% of the time. So $H_0 : p = 0.5$ vs $H_A : p < 0.5$.

By the central limit theorem, even this sample average is approximately normal. So we could use the techniques for sample means to calculate this p-value.

```
xbar   <- mean(clouds)
s      <- sd(clouds)
n      <- length(clouds)
z      <- (xbar - 0.5) / (s / sqrt(n))
pvalue <- pnorm(z)
pvalue


[1] 0.01213
```

Exact Calculation

But we know the sampling distribution of $\hat{p}$ under H_0 exactly.

$$n\hat{p} = \sum_i Z_i \sim Binomial(n, 0.5)$$

So we can calculate how extreme our observed $n\hat{p} = 179$ out of $n = 403$ is using the binomial distribution.

```
nphat <- sum(clouds)
pbinom(q = nphat, size = n, prob = 0.5)

[1] 0.01413
```

This is fairly close to the p-value using the normal approximation 0.0121.

Formal Connection Between Hypothesis Tests and CI's

Critical Value z_α

- If the P-value is less than α we reject H_0.
- For a two sided test This requires computing $P(|Z| \geq z)$, for the observed test statistic z, and comparing it to α.
- Alternatively we can find the critical value z_α such that $P(|Z| \geq z_\alpha) = \alpha$ and check if $|z| > z_\alpha$.
- For a one-sided test we find z_α such that $P(Z > z_\alpha) = \alpha$ and check if $z > z_\alpha$.

Hypothesis tests and CI's

A level α two-sided test rejects a hypothesis $H_0 : \mu = \mu_0$ exactly when the value of μ_0 falls outside a $(1 - \alpha)$ confidence interval for μ.

For example, consider a two-sided test of the following hypotheses

$$H_0 : \mu = \mu_0$$
$$H_a : \mu \neq \mu_0$$

at the significance level $\alpha = .05$.

Assume the test statistic is z and $2P(Z > |z|) = 2P(Z > z) = p < \alpha$. Let z_α be the critical value for level α. Assume the population SD is σ_0.

Hypothesis tests and CI's

$$p < \alpha$$

$$\Updownarrow$$

$$z > z_\alpha \quad \text{or} \quad z < -z_\alpha$$

$$\Updownarrow$$

$$\frac{\bar{x} - \mu_0}{\sigma_0/\sqrt{n}} > z_\alpha \quad \text{or} \quad \frac{\bar{x} - \mu_0}{\sigma_0/\sqrt{n}} < -z_\alpha$$

$$\Updownarrow$$

$$\mu_0 < \bar{x} - z_\alpha \cdot \frac{\sigma_0}{\sqrt{n}} \quad \text{or} \quad \mu_0 > \bar{x} + z_\alpha \cdot \frac{\sigma_0}{\sqrt{n}}$$

$$\Updownarrow$$

$$\mu_0 \notin [\bar{x} - z_\alpha \cdot \frac{\sigma_0}{\sqrt{n}}, \bar{x} + z_\alpha \cdot \frac{\sigma_0}{\sqrt{n}}]$$

μ_0 is not in the α confidence interval if and only if the null hypothesis is rejected at the α level.

Hypothesis tests and CI's

- If μ_0 is a value inside the 95% confidence interval for μ, then this test will have a *p*-value greater than .05, and therefore will not reject H_0.
- If μ_0 is a value outside the 95% confidence interval for μ, then this test will have a *p*-value smaller than .05, and therefore will reject H_0.

End of class examples

What's wrong

1. A significance test rejected the null hypothesis that the sample mean is equal to 500.
2. A test preparation company wants to test that the average score of its students on the ACT is better than the national average score of 21.2. The company states its null hypothesis to be $H_0 : \mu > 21.2$.
3. A study summary says that the results are statistically signficant and the p-value is 0.98.
4. The z test statistic is equal to 0.018. Because this is less than $\alpha = 0.05$, the null hypothesis was rejected.

Example

Sonnets by a certain Elizabethan poet are known to contain an average $\mu = 8.9$ new words (words not found in the poet's other works). The standard deviation of the number of new words is $\sigma = 2.5$. A new manuscript with six new sonnets has come to light and scholars are debating wheter it is the poet's work. The new sonnets contain an average of $\bar{x} = 10.2$ words not used in the poet's known works. We expect poems by another authro to contain more new words. Set up a hypothesis test, calculate a p-value, and form a conclusion.