# Tutorial5

Swarag

2023-09-24

# Hypothesis Testing vs Estimation

- Uses sample data to make inferences about the population from which the sample was taken.
- Estimates: puts bounds on the value of a population parameter
- Hypothesis Testing: asks only whether the parameter differs from a specific "null" expectation
- Estimates: How large is the effect?
- Hypothesis Testing: Is there any effect at all?

# Case study: Polio Vaccine, is it effective? IS THERE ANY EFFECT AT ALL?

Polio vaccine was tested on elementary-school students across the United States and Canada. In the study, 401,974 students were divided randomly into two groups: kids in one group received the vaccine, whereas those in the other group (the control group) were injected with saline solution instead. The students were unaware of which group they were in. Of those who received the vaccine, 0.016% developed paralytic polio during the study, whereas 0.057% of the control group developed the disease (Brownlee 1955). The vaccine seemed to reduce the rate of disease by two-thirds, but the difference between groups was quite small, only about four cases per 10,000. Did the vaccine work, or did such a small difference arise purely by chance?

# Let's do hypothesis testing

- Null Hypothesis($H_0$): The vaccine didn't work, and that any observed difference between groups happened only by chance
- Alternate Hypothesis($H_A$): The vaccine worked, and that the observed difference cannot be explained by chance events.

# What we do in hypothesis testing? How big of a difference is “SIGNIFICANT?”

- Assume that the null hypothesis($H_0$) is true(here: the vaccine has no effect).
- Calculate the probability of getting a difference between groups as big or bigger than that observed(called the p values)

For the vaccine, this probability turned out to be very small. Even though the rate of disease was not hugely different between the vaccine and control groups, the Salk vaccine trial was so large (over 400,000 participants) that it was able to demonstrate a real difference. Thus, the “null” hypothesis was rejected. The vaccine had an effect, sparing many kids from disease, which was borne out by the success of the vaccine in the ensuing decades.

# Summary of Concept

- Hypothesis testing quantifies how unusual the data are, assuming that the null hypothesis is true
- If the data are too different from what is expected by the null hypothesis, then we reject the null hypothesis.

*Hypothesis testing:*compares data to what we would expect to see if a specific null hypothesis were true. If the data are too unusual, compared to what we would expect to see if the null hypothesis were true, then the null hypothesis is rejected.

# Formalizing A Statistical Hypothesis

- Begins with clear statements of two hypotheses—the null and alternative hypotheses—about a population
- Both statistical hypotheses, the null and the alternative, are simple statements about a population

# The Null Hypothesis($H_0$)

- Specific claim about the value of a population parameter.
- Made for the purposes of argument and often embodies the skeptical point of view.
- A good null hypothesis is a statement that would be interesting to "reject".

# The Alternate Hypothesis($H_A$)

- Every null hypothesis is paired with an alternative hypothesis.
- Usually represents all other feasible parameter values except that stated in the null hypothesis.
- The alternative hypothesis typically includes possibilities that are experimentally more interesting than that stated in the null hypothesis.

# Decision: To reject or not to reject

- The null hypothesis is the only statement being tested with the data.
- If the data are consistent with the null hypothesis, then we say we have failed to reject it (we never "accept" the null hypothesis!).
- f the data are inconsistent with the null hypothesis, we reject it and say the data support the alternative hypothesis.
- Rejecting $H_0$ means that we have ruled out the null hypothesized value.
- Tells us in which direction the true value likely lies, compared to the null hypothesized value.
- This doesn't tell us about the magnitude.

# Summary: Basic Steps In Hypothesis Testing

1. State the hypotheses.
2. Compute the test statistic.
3. Determine the P-value.
4. Draw the appropriate conclusions.

# Example: The right hand of toad

Bisazza et al. (1996) tested the possibility of handedness in European toads, Bufo bufo, by sampling and measuring 18 toads from the wild. We will assume that this is a “random sample”. The toads were brought to the lab and subjected one at a time to the same indignity: a balloon was wrapped around each individual’s head. The researchers then recorded which forelimb each toad used to remove the balloon. It was found that individual toads tended to use one forelimb more than the other. At this point the question became: do right-handed and left-handed toads occur with equal frequency in the toad population. Of the 18 toads tested, 14 were right-handed and four were left-handed. Are these results evidence of a predominance of one type of handedness in toads?

# Example 1: The right hand of toad - stating the hypothesis

The number of interest is the proportion of right-handed toads in the population. Let's call this proportion p.

- $H_0$: Left- and right-handed toads are equally frequent in the population (i.e., $p = 0.5$).
- $H_A$: Left- and right-handed toads are not equally frequent in the population (i.e., $p \neq 0.5$).

This is a two-sided test. In a two-sided (or two-tailed) test, the alternative hypothesis includes parameter values on both sides of the parameter value specified by the null hypothesis.

# Example 1: Test statistic

- *Number calculated from the data that is used to evaluate how compatible the results are with those expected under the null hypothesis*
- For the toad study, we use the observed number of right-handed toads as our test statistic.
- Under the assumptions of null hypothesis, we are expected to see $\frac{9}{18}$ right handed frogs.
- But, we observed 14 $\implies$ 14 is our test statistics.

# Example 1: Null distribution

- Data doesn't always reflect the "truth", since there can be effects due to chance.
- Due to chance, there could be discrepancy between the observation and the null hypothesis even when the null hypothesis holds.
- This discrepancy could be quite large, event when the null hypothesis holds, specially when the number of observations are low.
- **Null Distribution:** sampling distribution of outcomes for a test statistic under the assumption that the null hypothesis is true.

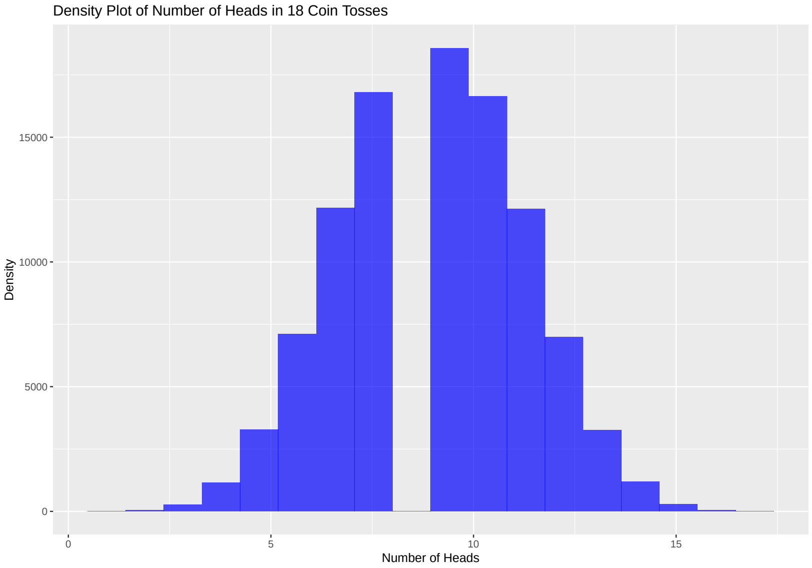
# Example 1: Null distribution- why we use it?

- To decide whether the data are compatible with the null hypothesis, we must calculate the probability of a mismatch as extreme as or more extreme than that observed, assuming that the null hypothesis is true.
- To obtain this probability, we need to determine the sampling distribution of the test statistic assuming that the null hypothesis is true.
- We need to determine what values of the test statistic are possible under $H_0$ and their associated probabilities.

# Example 1: Null distribution- how to compute it?

- The tricky part is to figure out what the null distribution is for the test statistic.
- Sampling 18 toads under H0 is like tossing 18 coins into the air and counting the number of heads that turn up when they land (letting heads represent right-handed toads)
- Tossing coins mimics well the sampling process under this $H_0$ because the probability of obtaining heads in any one toss is 0.5, which matches the null hypothesis.

# Example 1: Null distribution- visualization




visualization

# Example 1: Null distribution- probability

Now we need to see the probability of getting the values as larger or more larger than 14, assuming the null hypothesis is true(i.e, here the data generating process is consistent with null hypothesis).

```
pval = pbinom(4,18,0.5)+(1 - pbinom(14,18,0.5))
print(paste('Probability:',pval))
```

```
## [1] "Probability: 0.019210815429688"
```

Fourteen right-handed toads out of 18 total is not a perfect match to the expectation of the null hypothesis, but is the mismatch large enough to reject the possibility that chance alone is responsible?

# Example 1: Quantifying uncertainty: the P-value

- We want to know the probability of all results as unusual as or more unusual than that exhibited by the data.
- If this probability is small, then the null hypothesis is inconsistent with the data and we would reject the null hypothesis in favor of the alternative hypothesis.
- If the probability is not small, then we do not have enough evidence to doubt the null hypothesis, and we would not reject it.

**P-value:** the probability of obtaining the data (or data showing as great or greater difference from the null hypothesis) if the null hypothesis were true.

# Example 1: Draw the appropriate conclusion

- But what value of P is small enough?
- By convention in most areas of research, the boundary between small and not-small P-values is 0.05.
- We conclude from these data that most of the toads in the population are right-handed.

**Significance level**($\alpha$) : is a probability used as a criterion for rejecting the null hypothesis. If the P-value is less than or equal to $\alpha$, then the null hypothesis is rejected. If the P-value is greater than $\alpha$, then the null hypothesis is not rejected.

# Errors in hypothesis testing

- Rejecting $H_0$ does not necessarily mean that the null hypothesis is false.
- Similarly, failing to reject $H_0$ does not necessarily mean that the null hypothesis is true.
- This is because chance affects samples, sometimes with large impact.

# Type I & Type II errors

**Type I error:** is rejecting a true null hypothesis. The significance level $\alpha$ sets the probability of committing a Type I error.

**Type II error:** is failing to reject a false null hypothesis.

# Type I & Type II errors

- The significance level, $\alpha$, gives us the probability of committing a Type I error.
- Reducing $\alpha$ reduces Type I error.
- But, it makes it difficult to reject the null hypothesis, increasing Type II error.

# Power and Type II error

**Power of a test:** is the probability that a random sample will lead to rejection of a false null hypothesis.

- All else being equal, a study is better if it has more power.
- Power is difficult to quantify, because the probability of rejecting a null hypothesis depends on how different the truth is from the null hypothesis.
- We can usually estimate the power of a study. A study has more power if the sample size is large, if the true discrepancy from the null hypothesis is large, or if the variability in the population is low.