STAT 511

Lecture 14: Introduction to Hypothesis Testing

Devore: Section 8.1

Prof. Michael Levine

March 4, 2019

What is statistical hypothesis?

- A statistical hypothesis is a claim about the value of a parameter(s) or about the form of a distribution as a whole.
- As an example, consider a normal distribution with the mean μ. Then, the statement $\mu = .75$ is a hypothesis.

Null and Alternative Hypotheses

- Usually, two contradictory hypotheses are under consideration. For example, we may have $\mu = .75$ and $\mu \neq .75$. Alternatively, for a probability of success of some binomial distribution, we may have $p \geq .10$ and $p \leq .10$.
    1. The *null hypothesis* H_0 is the one that is initially assumed to be true.
    2. The *alternative hypothesis* H_a is the assertion contrary to H_0.

- We reject the null hypothesis in favor of the alternative hypothesis if the sample evidence suggests so. If the sample does not contradict H_0, we continue to believe it is true.
- Thus, the two possible conclusions from a hypothesis-testing analysis are *reject* H_0 or *fail to reject* H_0.

An Example of the Test

- A test of hypothesis is a method for using sample data to decide whether the null hypothesis should be rejected.
- How exactly do we formulate a test? It depends on what our goals are...
- Consider a company that wants to introduce an expensive new product to its line-up of existing ones. Clearly, there has to be an extensive evidence in favor of this new product. If it is, for example, a new type of the lightbulb, we need to ensure that its average lifetime is much longer than the one for existing types before adopting it.

- A reasonable test would be to test $H_0 : \mu = a$ vs. $H_a : \mu > a$ where a is some predetermined threshold.
- Clearly, the alternatives $H_a : \mu < a$ or $H_0 : \mu \neq a$ are of no interest in this case.
- $H_a : \mu < a$ and $H_a : \mu > a$ are called *one-sided alternatives*; $H_0 : \mu \neq a$ is called a *two-sided alternative*.
- The value a that separates null hypothesis from an alternative is called a *null value*.

Testing Procedure

- A test procedure is a rule, based on sample data, for deciding whether H_0 should be rejected.
- The key issue will be the following: Suppose that H_0 is in fact true. Then how likely is it that a (random) sample at least as contradictory to this hypothesis as our sample would result? Consider the following two scenarios:
  - There is only a .1% chance (a probability of .001) of getting a sample at least as contradictory to H_0 as what we obtained assuming that H_0 is true.
  - There is a 25% chance (a probability of .25) of getting a sample at least as contradictory to H_0 as what we obtained when H_0 is true

Testing procedure

- In the first scenario, something as extreme as our sample is very unlikely to have occurred when H_0 is true
- In the long run only 1 in 1000 samples would be at least as contradictory to the null hypothesis as the one we ended up selecting
- In contrast, for the second scenario, in the long run 25 out of every 100 samples would be at least as contradictory to H_0 as what we obtained assuming that the null hypothesis is true. So our sample is quite consistent with H_0, and there is no reason to reject it.

Example

- The company that manufactures brand D Greek-style yogurt is anxious to increase its market share
- They want to persuade those who currently prefer brand C to switch brands.
- The marketing department has devised the following blind taste experiment. Each of 100 brand C consumers will be asked to taste yogurt from two bowls, one containing brand C and the other brand D, and then say which one he or she prefers.
- The bowls are marked with a code so that the experimenters know which bowl contains which yogurt, but the experimental subjects do not have this information

Example

- Let p denote the proportion of all brand C consumers who would prefer C to D in such circumstances. Let us consider testing the hypotheses H_0: $p = .5$ versus H_0: $p < .5$
- The alternative hypothesis says that a majority of brand C consumers actually prefer brand D. Of course the brand D company would like to have H_0 rejected so that H_a is judged the more plausible hypothesis.
- If the null hypothesis is true, then whether a single randomly selected brand C consumer prefers C or D is analogous to the result of flipping a fair coin.

Example

- Let X = the number among the 100 selected individuals who prefer C to D. This random variable will serve as our test statistic, the function of sample data on which we will base our conclusion.
- Now X is a binomial random variable (the number of successes in an experiment with a fixed number of independent trials having constant success probability p). When H_0 is true, this test statistic has a binomial distribution with $p = .5$, in which case $E(X) = np = 100(.5) = 50$

Example

- Intuitively, a value of X "considerably" smaller than 50 argues for rejection of H_0. Intuitively, a value of X "considerably" smaller than 50 argues for rejection of H_0 in favor of H_0
- Suppose the observed value of X is $x = 37$. How contradictory is this value to the null hypothesis? To answer this question, let us first identify values of X that are even more contradictory to H_0 than is 37 itself.
- Clearly 35 is one such value, and 30 is another; in fact, any number smaller than 37 is a value of X more contradictory to the null hypothesis than is the value we actually observed.

Example

- Now consider the probability, computed assuming that the null hypothesis is true, of obtaining a value of X at least as contradictory to H_0 as is our observed value:
- $$P(X \leq 37 |\ H_0 \text{ is true }) =$$
$$P(X \leq 37 | X \sim Bin(100, .5)) = B(37; 100, .5) = .006$$
- Thus if the null hypothesis is true, there is less than a 1% chance of seeing 37 or fewer successes among the 100 trials. This suggests that $x = 37$ is much more consistent with the alternative hypothesis than with the null, and that rejection of H_0 in favor of H_a is a sensible conclusion.

Example

- In addition, note that $\sigma_x = \sqrt{npq} = \sqrt{100(.5)(.5)} = 5$ when H_0 is true. It follows that 37 is more than 2.5 standard deviations smaller than what we would expect to see were H_0 true.
- Now suppose that 45 of the 100 individuals in the experiment prefer C (45 successes). Let us again calculate the probability, assuming H_0 true, of getting a test statistic value at least as contradictory to H_0 as this:

$$P(X \leq 45 |\ H_0 \text{ is true }) = P(X \leq 45 | X \sim Bin(100, .5)) = B(45; 100, .5) = .184$$

Example

- So if in fact $p = .5$, it would not be surprising to see 45 or fewer successes.
- For this reason, the value 45 does not seem very contradictory to H_0 (it is only one standard deviation smaller than what we would expect were H_0 true). Rejection of H_0 in this case does not seem sensible.

Testing procedures

- A test statistic is a function of sample data used as basis for deciding whether H_0 should be rejected. The selected test statistic should discriminate effectively between the two hypotheses. That is, values of the statistics that result when H_0 is true should be quite different from those that result when H_a is true
- The P-value is the probability, calculated assuming that the null hypothesis is true, of obtaining a value of test statistic at least as contradictory to H_0 as the value calculated from the available sample data. A conclusion is reached in a hypothesis testing analysis by selecting a number α, (the level of significance) of the test, that is reasonably close to zero. Then, H_0 will be rejected in favor H_a if P-value $\leq \alpha$.

Testing procedures

- Common values of α are 0.1,, 0.05, and 0.01
- For example, if we select a significance level of .05 and then compute P-value $= .0032$, H_0 would be rejected because $.0032 \leq .05$
- With this same P-value, the null hypothesis would also be rejected at the smaller significance level of .01 because $.0032 \leq .01$. However, at a significance level of .001 we would not be able to reject H_0 since $.0032 \geq .001$.

So what is a P-value?!

- The P-value is a probability
- This probability is calculated assuming that the null hypothesis is true
- To determine the P-value, we must first decide which values of the test statistics are at least as contradictory to H_0 as the value obtained from our sample
- The smaller the P-value, the stronger is the evidence against H_0 and in favor of H_a
- The P-value is not the probability that the null hypothesis is true, nor is it the probability that the erroneous conclusion has been reached

Errors in Hypotheses testing

- A type I error consists of rejecting H_0 when it is true
- A type II error consists of not rejecting H_0 when it is false
- Example: a cereal manufacturer claims that a serving of one of its brands provides 100 calories
- Interesting info: calorie content used to be determined by a destructive testing method, but the requirement that nutritional information appear on packages has led to more straightforward techniques
- Of course the actual calorie content will vary somewhat from serving to serving (of the specified size)
- Thus, 100 should be interpreted as an average. It could be distressing to consumers of this cereal if the true average calorie content exceeded the asserted value

Errors in Hypotheses testing

- An appropriate formulation of hypotheses is to test H_0: $\mu = 100$ versus H_0: $\mu > 100$.
- The alternative hypothesis says that consumers are ingesting on average a greater amount of calories than what the company claims
- A type I error here consists of rejecting the manufacturer's claim that $\mu = 100$ when it is actually true. A type II error results from not rejecting the manufacturer's claim when it is actually the case that $\mu > 100$.
- The only way to get rid of both errors is to use the entire population! In reality, a different procedure has to be followed.

Example

- Assume that 25% of the time automobiles have no visible damage in 10mph crash tests. Denote p the proportion of all 10 mph crashes that results in no visible damage to the new bumper. Then, $H_0 : p = .25$ vs. $H_a : p > .25$. The experiment is based on $n = 20$ independent crashes with prototype of the new design.

Type I Error analysis

1. Consider the following procedure:
   1.1 Test Statistic is X - the number of crashes with no visible damage
   1.2 If H_0 is true, $E(X) = np_0 = (20)(0.25) = 5$. Intuition suggests that an observed value x much larger than this would provide strong evidence against H_0 and in support of H_a
   1.3 Consider using a significance level of .10. The P-value is $P(X \geq x | X \sim Bin(20, .25) = 1 - B(x-1; 20, .25)$ for $x > 0$.
2. Check that $P(X \geq 7) = .214$ and $P(X \geq 8) \approx .10$, $P(X \geq 9) = 0.041$
3. Rejecting H_0 when P-value $\leq .10$ is equivalent to rejecting H_0 when $X \geq 8$
4. Thus, the probability of Type I error is

$$\alpha = P(\text{ Type I Error }) = P(X \geq 8 \text{ when } X \sim Bin(20, .25))$$
$$= 1 - B(7; 20, .25) = .102$$

Type I Error Analysis

- That is, the probability of type I error is just the significance level α
- If the null hypothesis is true here and the test procedure is used over and over again, each time in conjunction with a group of 20 crashes, in the long run the null hypothesis will be incorrectly rejected in favor of the alternative hypothesis about 10% of the time.
- So our test procedure offers reasonably good protection against committing a type I error

Type II Error Analysis

- There is only one type I error probability because there is only one value of the parameter for which H_0 is true (this is one benefit of simplifying the null hypothesis to a claim of equality).
- Let β denote the probability of committing a type II error. Unfortunately there is not a single value of β, because there are a multitude of ways for H_0 to be false it could be false because p = .30, because p = .37, because p = .5, and so on.
- There is in fact a different value of β for each different value of p that exceeds .25

Type II Error Analysis

- Suppose the true value of p is $p = 0.3$. Then,

$$\begin{aligned}\beta(.3) &= P(\text{ Type II Error when } p = 0.3)\\ &= P(X \leq 7 \text{ when } X \sim Bin(20, .3))\\ &= B(7; 20, .3) = .772\end{aligned}$$

- It is easy to understand that β decreases as p grows more different from the null value .25

Type II Error Analysis

- The accompanying table displays β for selected values of p (each calculated as we just did for $\beta(.3)$). Clearly, β decreases as the value of p moves farther to the right of the null value .25. Intuitively, the greater the departure from H_0, the more likely it is that such a departure will be detected.

p	.3	.4	.5	.6	.7	.8
$\beta(p)$	.772	.416	.132	.021	.001	.000

- The probability of committing a type II error here is quite large when $p = .3$ or .4. This is because those values are quite close to what H_0 asserts and the sample size of 20 is too small to permit accurate discrimination between .25 and those values of p.

Example

- The proposed test procedure is still reasonable for testing the more realistic null hypothesis that $p \leq .25$. In this case, there is no longer a single type I error probability α,, but instead there is an α for each p that is at most .25: $\alpha(.25)$, $\alpha(.23)$, $\alpha(.20)$, $\alpha(.15)$, and so on.
- It is easily verified, though, that $\alpha(p) < \alpha(.25) = .102$ if $p < .25$. That is, the largest type I error probability occurs for the boundary value .25 between H_0 and H_1.
- Thus if α is small for the simplified null hypothesis, it will also be as small as or smaller for the more realistic H_0

Errors in test procedure

- It is no accident that in the two foregoing examples, the significance level α turned out to be the probability of a type I error.
- Proposition: the test procedure that rejects H_0 if P-value $\leq \alpha$ and otherwise does not reject H_0 has the level of significance has P(Type I error)$= \alpha$

Errors in test procedure

- The inverse relationship between the significance level α and type II error probabilities in Example 8.5 can be generalized in the following manner:
- Proposition: suppose an experiment or a sampling procedure is selected, a sample size is specified, and a test statistic is chosen. Then, increasing the level of significance α, i.e. employing the larger Type I error probability, results in a smaller value of β for any particular parameter value consistent with H_a
- This result is intuitively obvious because when α is increased, it becomes more likely that we'll have P-value $\leq \alpha$ and therefore less likely that P-value $> \alpha$

Errors in test procedure

- This proposition implies that once the test statistic and n are fixed, it is not possible to make both α and any values of β that might be of interest arbitrarily small.
- Deciding on an appropriate significance level involves compromising between small α and small β. In Example 8.5, the type II error probability for a test with $\alpha = .01$ was quite large for a value of α close to the value in H_0.
- A strategy that is sometimes (but perhaps not often enough) used in practice is to specify α and also β for some alternative value of the parameter that is of particular importance to the investigator.

Errors in test procedure

- In practice it is usually the case that the hypotheses of interest can be formulated so that a type I error is more serious than a type II error.
- The approach adhered to by most statistical practitioners is to reflect on the relative seriousness of a type I error compared to a type II error and then use the largest value of α that can be tolerated.
- This amounts to doing the best we can with respect to type II error probabilities while ensuring that the type I error probability is sufficiently small.

Errors in test procedure

- For example, if $\alpha = .05$ is the largest significance level that can be tolerated, it would be better to use that rather than $\alpha = .01$, because all β for the former α will be smaller than those for the latter one.
- As previously mentioned, the most frequently employed significance levels are $\alpha = .05$, .01, .001, and .10.

Exceptions

- Here is one example from particle physics: according to the article "Discovery or Fluke: Statistics in Particle Physics" (Physics Today, July 2012: 4550), "the usual choice of alpha is 3×10^{-7}, corresponding to the 5σ of a Gaussian [i.e., normal] H_0 distribution. "
- Why so stringent? For one thing, recent history offers many cautionary examples of exciting 3σ and 4σ signals that went away when more data arrived