Chapter 8. Statistical Inference

8.3: Introduction to Hypothesis Testing

(From "Probability & Statistics with Applications to Computing" by Alex Tsun)

Hypothesis testing allows us to "statistically prove" claims. For example, if a drug company wants to claim that their new drug reduces the risk of cancer, they might perform a hypothesis test. Or if a company wanted to argue that their academic prep program leads to a higher SAT score. A lot of business decisions are reliant on this statistical method of hypothesis testing, and we'll see how to conduct them properly below.

8.3.1 Hypothesis Testing (Idea)

Suppose we have this Magician Mark, who says

Magician Mark: I have here a fair coin.

And then an audience member, a skeptical statistician named Stacy, engages him in a conversation:

Skeptical Statistician Stacy: I don't believe you. Can we examine it?
Magician Mark: Be my guest.
Skeptical Statistician Stacy: I'm going to flip your coin 100 times and see how many heads I get.
[Stacy flips the coin 100 times and sees 99 heads.]
You cannot be telling the truth, there's no way this coin is fair!
Magician Mark: Wait I was just unlucky, I swear I'm not lying!

So let's give Mark the **benefit of the doubt**. We'll compute the probability that we observed an outcome *at least as extreme* as this, **given that Mark isn't lying**.

If Mark isn't lying, then the coin is fair, so the number of heads observed should be $X \sim \text{Bin}(100, 0.5)$, because there are 100 independent trials and a 50% of heads since it's fair. So, the probability that we observe at least 99 heads (because we're looking for something *as least as extreme*), is the sum of the probability of 99 heads and the probability of 100 heads. You just sum the Binomial PMF and you get:

$$\mathbb{P}(X \geq 99) = \binom{100}{99}(0.5)^{99}(1-0.5)^1 + \binom{100}{100}(0.5)^{100} = \frac{101}{2^{100}} \approx 7.96 \times 10^{-29} \approx 0$$

Basically, if the coin were fair, the probability of what we just observed (99 heads or more) is basically 0. This is strong statistical evidence that the coin is NOT fair. Our assumption was that the coin is fair, but if this were the case, observing such an extreme outcome would be extremely unlikely. Hence, our assumption is probably wrong.

So, this is like a "Probabilistic Proof by Contradiction"!

8.3.2 Hypothesis Testing (Example)

There is a formal procedure for a hypothesis test, which we will illustrate by example. There are many types of hypothesis tests, each with different uses, but we'll get into that later! You'll see the CLT often appear in the most fundamental/commonly conducted hypothesis tests.

1. **Make a claim (like "Airplane food is good", "Pineapples belong on pizza", etc...)**
   - Our example will be that SuperSAT Prep claims that their program helps students perform better on the SAT. (The average SAT score as of June 2020 was: 1059 out of 1600, and the standard deviation of SAT scores was 210).
2. **Set up a null hypothesis H_0 and alternative hypothesis H_A.**
   (a) Alternative hypothesis can be one-sided or two-sided.
   - Let μ be the true mean of the SAT scores of students of SuperSAT Prep.
   - Our **null hypothesis** is that $H_0 : \mu = 1059$, which is our "baseline", "no effect", "benefit of the doubt". We're going to assume that the true mean of our scores is the same as the nationwide scores (for the sake of contradiction).
   - Our **alternative hypothesis** is what we want to show, which is $H_A : \mu > 1059$, or that SuperSAT Prep is good and that their test takers are (strictly) better off. So, our alternative will assert that $\mu > 1059$.
   - This is called a **one-sided hypothesis**. The other one-sided hypothesis would be $\mu < 1059$ (if we wanted to argue that SuperSAT Prep makes students worse off).
   - A **two-sided hypothesis** would be that $\mu \neq 1059$, because it's two sides (less than or greater than). This is if we wanted to argue that SuperSAT Prep makes some difference for better or worse.
3. **Choose a significance level α (usually $\alpha = 0.05$ or 0.01).**
   - Let's choose $\alpha = 0.05$ and explain this more later!
4. **Collect data.**
   - We observe 100 students from SuperSAT Prep, $x_1, ..., x_{100}$. It turns out, the sample mean of the scores, $\bar{x}$, is $\bar{x} = 1113$.
5. **Compute a p-value, $p = \mathbb{P}$(observing data at least as extreme as ours $\mid H_0$ is true).**
   - Again, since we're assuming H_0 is true (that SuperSAT has no effect), our true mean μ is 1059 (again we do this in hopes of reaching a "probabilistic contradiction"). By the CLT, since $n = 100$ is large, the distribution of the sample mean of 100 samples is approximately normal with mean 1059, and variance $\frac{210^2}{100}$ (because the variance of a single test taker was given to be $\sigma^2 = 210^2$, and so the variance of the sample mean is $\frac{\sigma^2}{n}$):

$$\bar{X} \approx \mathcal{N}(\mu = 1059, \sigma^2 = \frac{210^2}{100})$$

   So, then, the p-value is the probability that if we took an arbitrary sample mean, that it would be at least as extreme as the one we computed, which was 1113. So, we can just standardize, look up a Φ table like always, which is a procedure you know how to do:

$$p = \mathbb{P}\left(\bar{X} \geq \bar{x}\right) = \mathbb{P}\left(\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \geq \frac{\bar{x} - \mu}{\sigma/\sqrt{n}}\right) = \mathbb{P}\left(Z \geq \frac{1113 - 1059}{210/\sqrt{100}}\right) = \mathbb{P}\left(Z \geq 2.14\right) \approx 0.0162$$

We end up getting that our p-value is 0.0162

6. **State your conclusion. Include an interpretation in the context of the problem.**

   (a) If $p < \alpha$, "reject" the null hypothesis H_0 in favor of the alternative H_A. (Because, given the null hypothesis is true, the probability of what we saw happening (or something more extreme) is p which is less than some small number α.)

   (b) Otherwise, "fail to reject" the null hypothesis H_0.

   - Since $p = 0.0162 < 0.05 = \alpha$, we'll reject the null hypothesis H_0 at the $\alpha = 0.05$ significance level. We can say that there is strong statistical evidence to suggest that SuperSAT Prep actually helps students perform better on the SAT.

     Notice that if we had chosen $\alpha = 0.01$ earlier instead of 0.05, we would have a different conclusion: Since $p = 0.0162 > 0.01 = \alpha$, we fail to reject the null hypothesis at the $\alpha = 0.01$ significance level. There is insufficient evidence to prove that SuperSAT Prep actually helps students perform better.

     Note that **we'll NEVER say we "accept" the null hypothesis.** If you recall the coin example, if we had observed 55 heads instead of 99, that wouldn't have been improbable. We wouldn't have called the magician a liar, but it does NOT imply that $p = 0.5$. It could have been 0.54 or 0.58, for example.

8.3.3 Hypothesis Testing Procedure

The formal hypothesis testing procedure is summarized as follows:

1. Make a claim (like "Airplane food is good", "Pineapples belong on pizza", etc...)
2. Set up a null hypothesis H_0 and alternative hypothesis H_A.
   (a) Alternative hypothesis can be one-sided or two-sided.
   (b) The null hypothesis is usually a "baseline", "no effect", or "benefit of the doubt".
   (c) The alternative is what you want to "prove", and is opposite the null.
3. Choose a significance level α (usually $\alpha = 0.05$ or 0.01).
4. Collect data.
5. Compute a p-value, $p = \mathbb{P}\left(\text{observing data at least as extreme as ours} \mid H_0 \text{ is true}\right)$.
6. State your conclusion. Include an interpretation in the context of the problem.
   (a) If $p < \alpha$, "reject" the null hypothesis H_0 in favor of the alternative H_A. We say our result is **statistically significant** in this case!
   (b) Otherwise, "fail to reject" the null hypothesis H_0.

8.3.4 Exercises

1. You want to determine whether or not more than 3/4 of Americans would vote for George Washington for President in 2020 (if he were still alive). In a random poll sampling $n = 137$ Americans, we collected responses $x_1, \ldots, x_n$ (each is 1 or 0, if they would vote for him or not). We observe 131 "yes" responses: $\sum_{i=1}^{n} x_i = 131$. Perform a hypothesis test and state your conclusion.

   Solution: We have our claim that "Over 3/4 of Americans would vote for George Washington for President in 2020 (if he were still alive).

   Let p denote the true proportion of Americans that would vote for Washington. Then our null and alternative hypotheses are:

   $$H_0 : p = 0.75$$
   $$H_A : p > 0.75$$

   Let's test these hypotheses at the $\alpha = 0.01$ significance level.

   We know by the CLT that the sample mean is approximately $\bar{X} \sim \mathcal{N}\left(\mu = 0.75, \sigma^2 = \frac{0.75(1-0.75)}{137}\right) = \mathcal{N}(0.75, \sigma^2 = 0.037^2)$ (since $X_i \sim \text{Ber}(p)$: $\mathbb{E}[X_i] = p = 0.75$ under the null hypothesis, and $\mathsf{Var}(X_i) = p(1-p) = 0.75(1-0.75)$ and we know $\mathbb{E}\left[\bar{X}\right] = \mathbb{E}[X_i] = p$ and $\mathsf{Var}\left(\bar{X}\right) = \frac{\sigma^2}{n} = \frac{0.75(1-0.75)}{n}$).

   Hence our p-value (observing data at least as extreme), is

   $$\mathbb{P}\left(\bar{X} \geq \bar{x}\right) = \mathbb{P}\left(\mathcal{N}(0.75, \sigma^2 = 0.037^2) \geq \frac{131}{137}\right) = \mathbb{P}\left(Z \geq \frac{131/137 - 0.75}{0.037}\right) = \mathbb{P}\left(Z \geq 5.42643\right) \approx 0$$

   With a p-value so close to 0 (and certainly $< \alpha = 0.01$), we reject the null hypothesis that (only) 75% of Americans would vote for Washington. There is strong evidence that this proportion is actually larger.

   Note: Again, what we did was: assume $p = 0.75$ (null hypothesis), then note that the probability of observing data so extreme (in fact very close to 100% of people), was nearly 0. Hence, we reject this null hypothesis because what we observed would've been so unlikely if it were true.