Notes: Hypothesis Testing, Fisher's Exact Test

CS 3130 / ECE 3530: Probability and Statistics for Engineers

March 28, 2024

The Lady Tasting Tea

Many of the modern principles used today for designing experiments and testing hypotheses were introduced by Ronald A. Fisher in his 1935 book *The Design of Experiments*. As the story goes, he came up with these ideas at a party where a woman claimed to be able to tell if a tea was prepared with milk added to the cup first or with milk added after the tea was poured. Fisher designed an experiment where the lady was presented with 8 cups of tea, 4 with milk first, 4 with tea first, in random order. She then tasted each cup and reported which four she thought had milk added first. Now the question Fisher asked is, "how do we test whether she really is skilled at this or if she's just guessing?"

To do this, Fisher introduced the idea of a **null hypothesis**, which can be thought of as a "default position" or "the status quo" where nothing very interesting is happening. In the lady tasting tea experiment, the null hypothesis was that the lady could not really tell the difference between teas, and she is just guessing. Now, the idea of hypothesis testing is to attempt to *disprove* or *reject* the null hypothesis, or more accurately, to see how much the data collected in the experiment provides evidence that the null hypothesis is false.

The idea is to *assume* the null hypothesis is true, i.e., that the lady is just guessing. Under this assumption and given the outcome of the experiment, we can now compute the probability of her performing as well as she did or better. Let's see how this works with an example outcome. Let's assume the lady gets all 8 cups correct. We can build a table of this outcome (this is called a **contingency table**):

		Lady's Answer	
		Milk First	Tea First
Truth	Milk First	4	0
	Tea First	0	4

Under the null hypothesis assumption (that she is guessing), what is the probability of this outcome? The lady knows there are exactly 4 cups of each, so she is essentially choosing 4 cups at random out of 8. There are "8 choose 4" ways to do this, so her probability is

$$P(\text{"all correct"}) = \frac{1}{\text{"number of ways to guess"}} = \frac{1}{\binom{8}{4}} = \frac{1}{70} \approx 0.014.$$

So, if she is guessing, there is only a 1.4% chance that she will get all cups correct. Now, let's look at the general situation. Notice that if we set how many cups with milk first that she gets correct, this determines the entire table. This is because she knows to choose 4 cups in each category, and thus each row and each column must sum to 4. The table for the general outcome looks like this:

		Lady's Answer	
		Milk First	Tea First
Truth	Milk First	k	4 - k
	Tea First	4 - k	k

First, notice that correct answers are on the diagonal. So, a value of k means that the lady actually has $2k$ answers correct. Here the counting problem to compute the probability of a general outcome is more difficult, but it follows what is called a **hypergeometric distribution**. (We won't cover this distribution in detail, but see the Wikipedia article if you want to learn more about it.) The probability becomes:

$$p(k) = \frac{\binom{4}{k}\binom{4}{4-k}}{\binom{8}{4}} = \frac{1}{70}\binom{4}{k}^2.$$

Now, this is the probability of the lady getting *exactly* $2k$ answers correct. What we originally wanted to ask is "what is the probability of her getting this outcome *or better*?" To get this, we need to sum over all values k or greater (up to the max of 4). Letting X be the total number of correct answers, this is:

$$P(\text{“}2k \text{ correct or better”}) = P(X \geq 2k) = \sum_{i=k}^{4} p(i).$$

Here are the probabilities of the 5 possible outcomes for the experiment:

$$p(0) = \frac{1}{70}, \quad p(1) = \frac{16}{70}, \quad p(2) = \frac{36}{70}, \quad p(3) = \frac{16}{70}, \quad p(4) = \frac{1}{70}.$$

Notice the symmetric in the probabilities. It is just as hard to get all wrong as it is to get all correct! Finally, the probabilities for getting $2k$ correct answers or better are

$$P(X \geq 0) = 1, \quad P(X \geq 2) = \frac{69}{70}, \quad P(X \geq 4) = \frac{53}{70}, \quad P(X \geq 6) = \frac{17}{70}, \quad P(X \geq 8) = \frac{1}{70}.$$

By the way, according to the legend, the lady got all 8 cups correct!

Summary of General Hypothesis Test Procedure:

1. Define the **null hypothesis**, which is the uninteresting or default explanation.
2. Assume that the null hypothesis is true, and determine the probability rules for the possible outcomes of the experiment.
3. After collecting data, compute the probability of the final outcome or even more extreme outcomes.

Further reading:
Ronald Fisher: `http://en.wikipedia.org/wiki/Ronald_A._Fisher`
Fisher's Exact Test: `http://en.wikipedia.org/wiki/Fisher's_exact_test`
Hypergeometric Distribution: `http://en.wikipedia.org/wiki/Hypergeometric_distribution`

Hypothesis Testing Procedure
We will now formalize this **hypothesis testing** procedure so we can use it more generally.

Step 1: hypothesis formulation.
This is to analyze a process, or an algorithm, or an ongoing natural phenomenon. It does not start with data, the data will come later. Like the lady's ability to detect milk or tea first. The first step is to make a hypothesis about that algorithm; often we actually ask for two hypotheses. The **null hypothesis** H_0 is a bland or boring hypothesis (e.g., the lady will guess at random). This one will need to be more precise, and will require a probability distribution. The **alternative hypothesis** H_1 can be a bit less precise, but is where the real conjecture lies; it needs to specify a way that we think the real algorithm is distinct from the null hypothesis (that the lady can distinguish milk first from tea first).

Lets now consider another example. One where we think UU students are taller than the national average. We look up reports on heights, and model this as a normal distribution $N(67, 25)$, so mean of 67 inches and variance of 25 inches squared. This is our null distribution for heights of UU students. For our alternative hypothesis, we can assume the variance is the same, but assume that the mean $\mu_{UU} > 67$. Note that we do not need to specify the actually guess for μ_{UU} here.

Step 2: Design experiment.
We now want to design an experiment that will evaluate if we should deviate from the null hypothesis. So we typically take a random sample $X_1, X_2, \ldots, X_n$ and analyze a statistic $T(X_1, \ldots, X_n)$. This generically will be called a **test statistic** T. In the Utah height example, we use our usual sample mean statistic $\bar{X}_n$.

Next we need to decide on a confidence threshold for how much evidence we would need to deviate from our simple null hypothesis. This is formulated in terms of probability that our test statistic (or something more extreme) happens less than some fraction α of the time. A typical choice for α is 0.05 or 0.01, but there is no one right value. In particular, we define a **critical value at** α, that is a threshold t_α so that $Pr(T \leq t_\alpha) = 1 - \alpha$.

When $X_i \sim N(\mu, \sigma^2)$, then we have that the distribution of $\bar{X}_n$ is according to $N(\mu, \sigma^2/n)$ and that $Z_n = \frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \sim N(0, 1)$. In this case we can define $t_\alpha = \mu + z_\alpha \frac{\sigma}{\sqrt{n}}$, where z_α is the $(1 - \alpha)$-quantile of the normal distribution.

Step 3: Run Experiment.
Now finally, we can collect the data $x_1, x_2, \ldots, x_n$. These are the realizations of the random vari-

ables $X_1, \ldots, X_n$. We compute the realization of the test statistic $t = T(x_1, \ldots, x_n)$, and compare t to t_α.

If $t > t_\alpha$, then we can **reject the null hypothesis**. That is, we say there is sufficient evidence that an alternative hypothesis is sufficiently more likely. In particular, under our null hypothesis, the data we found (or less extreme data) is unlikely to occur under that hypothesis more than $(1-\alpha)100\%$ of the time. This does not mean the alternative hypothesis correct, but basically there is an alternative hypothesis of the form we consider that is sufficiently more likely than the one we chose as null, given the evidence we found.

If $t \leq t_\alpha$, then we **do not reject the null hypothesis**. That is, we did not find evidence that an alternative hypothesis, for the forms we consider, is significantly more likely (at a $(1-\alpha)100\%$ critical value) than the null hypothesis. This case *does not confirm the null hypothesis*, it just does not show sufficient evidence against it.

Sometimes we also compute a **p-value**. This is the value p such that $Pr(T \leq t) = 1 - p$. That is, it is the probability under the null hypothesis, that a test statistic T would be as larger or larger than the one we found t.

Modified example without known variance.

In the above height example, we assumed a null distribution of $N(67, 25)$. But what if we only have a null guess of the average height, not the standard deviation? That is if we only want to assume $N(67, \sigma^2)$ for some unknown σ^2, what should we do?
Use a t distribution/statistic instead of a normal one.

The test statistic is now $T = \frac{\bar{X}_n - \mu}{S_n/\sqrt{n}}$ where $\mu = 67$ inches, S_n is the sample standard deviation, and n is the number of samples we drew. We can then set the *critical value at* α as t_α, which recall is the $1-\alpha$ quantile for the t-distribution $t(n-1)$. Then on real data realizations $x_1, \ldots, x_n$ we can compute $\bar{x}_n$ and then S_n^2 and finally a realization of the test statistic as $t = \frac{\bar{x}_n - 67}{S_n/\sqrt{n}}$. With this we can compare to t_α to see if it is above or below the critical value at α. We can also compute its p-value as $1 - \int_{a=t}^{\infty} t(a, df = n-1)$ or in R as

```
p = 1-pt(t, df=n-1).
```

Paired sample t-test.

The paired sample t-test is used to determine whether the mean difference between two sets of observations *of the same subjects* is zero. In a paired sample t-test, each subject or object is measured twice, resulting in pairs of observations. Statistics are done on the *difference* of these observations (per sample), and statistics on done on the means of those differences. The null hypothesis is typically "no effect", in which case the mean of the differences would be zero.
This analysis would come up, for instance, in examining if/how a particular set of individuals responds to a treatment. Are their symptoms better or worse after treatment?
The procedure is typically as follows:

1. Subtract the two measurements for each individual – across, for instance, n subjects.
2. Compute the sample mean and variance, $\bar{X}_n$ and S_n^2 respectively.
3. Build the statistic $T = \frac{\bar{X}_n - \mu}{S_n/\sqrt{n}}$, where μ is the mean of the null hypothesis (zero if "no effect").

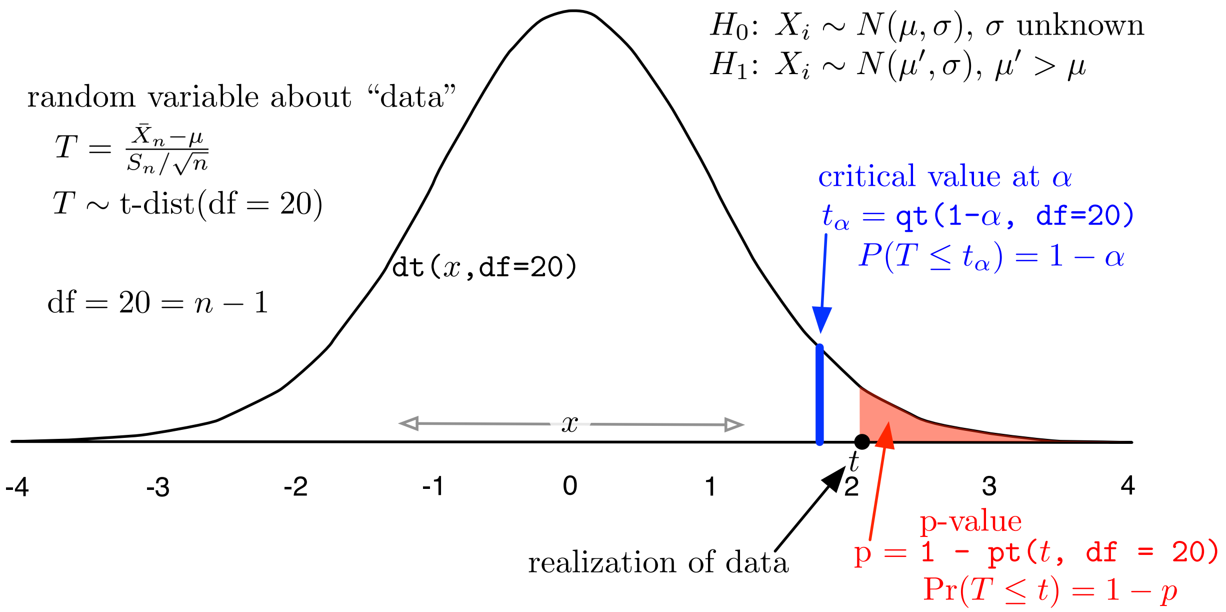
4. Identify a critical value for the alternative hypothesis (either left tailed, right tailed, or double tailed – depending on the alternative hypothesis).

5. Reject or not reject the null hypothesis based on the comparison of T to the critical value.

Below is an example of R code that runs this paired t-test on a set of data.

```
rawdata = c(1, 90.563, 110.642,
2, 94.816, 101.588,
3, 109.56, 120.607,
4, 90.222, 83.2217,
5, 97.598, 109.272,
6, 91.167, 115.806,
7, 96.65, 99.8958,
8, 97.616, 117.94,
9, 88.845, 106.052,
10, 90.817, 82.8229,
11, 89.294, 116.639,
12, 115.83, 128.61,
13, 121.29, 119.665,
14, 87.872, 108.383,
15, 93.793, 96.3738)

data = array(rawdata, dim = c(3,15))
print(data)
df = data.frame(t(data))
colnames(df) = c("ID", "Before", "After")
diff = df$After - df$Before
print(diff)
s = sqrt(var(diff))
x_d = mean(diff)
print("Here is the critical value")
print(qt(0.95, df = length(diff)-1))
stat = x_d/(s/sqrt(length(diff))) print("Here is our test statistic")
print(stat)
print("P-value is")
print(1- pt(stat, length(diff)-1))
```


H_0: $X_i \sim N(\mu, \sigma)$, σ unknown
H_1: $X_i \sim N(\mu', \sigma)$, $\mu' > \mu$
random variable about "data"
$T = \frac{\bar{X}_n - \mu}{S_n/\sqrt{n}}$
$T \sim$ t-dist(df $= 20$)
df $= 20 = n - 1$
dt(x,df=20)
critical value at α
$t_\alpha =$ qt(1-α, df=20)
$P(T \leq t_\alpha) = 1 - \alpha$
x
t
-4
-3
-2
-1
0
1
2
3
4
realization of data
p-value
p = 1 - pt(t, df = 20)
$\Pr(T \leq t) = 1 - p$