More Probability and Hypothesis Testing Review

Lizzy Huang
Duke University

Week 3

In Week 3, we will talk about decision making. Under the frequentist framework, we can only infer a parameter by doing hypothesis testings and calculating confidence intervals. We may make decision based on the results. However, in either way we cannot generate a "probabilistic" argument say "the probability for a parameter lying within an interval is 95%". In the Bayesian framework, what we focus on is now obtaining the posterior distribution of the parameter, instead of a single point estimate. Due to the fact that Bayesian approach does provide us the probability distribution, we may analysize the situation and make decision not only based on a single point, but also on the distribution, which accounts for the uncertainty. In this **Review**, we will introduce more basic probability concepts, as well as, review the Student's t-distribution and the frequentist hypothesis testing for means.

Expected Value (new)

In the frequentist framework, expected value of a random variable X, intuitively, is the long-run average value of repetitions of the experiment it represents. Mathematically, it is calculated as the probability-weighted average of all possible values.

In the Bayesian point of view, since we no longer consider repeating the same experiment in a long run, we adopt the mathematical calculation when we talk about the expected value of a random variable X. We denote the expected value to be $E(X)$.

When the random variable X is discrete, $E(X)$ is given by

$$E(X) = \sum_{\text{all } k} kP(X = k), \tag{1}$$

where k represents the value that X can take.

When the random variable X is continuous, $E(X)$ is given by

$$E(X) = \int_{-\infty}^{\infty} xp(x)\, dx. \tag{2}$$

No matter X is discrete or continuous, $E(X)$ is always of a form as the sum of value × weight, where the "weight" is given by the probability mass function or probability density function. **Expected value is also called the expected mean, the expected average, or just simply mean or average, according to the context.**

Example

> Suppose we have a bag of 5 M&M's. The general percentage of yellow M&M's in a bag is 10%, i.e., 0.1. Let X be the number of yellow M&M's in this bag. What is the expected number of yellow M&M's we would have?

Recall that this is a Binomial process, with 5 "trials" (5 M&M's in one bag), and "probability of success" $p = 0.1$. To get the expected number of yellow M&M's, we need to first obtain the general form of the probability mass function (pmf) for X. Since X follows a Binomial distribution, the probability for X to take each k ($k = 0, 1, 2, 3, 4, 5$) is

$$P(X = k) = \binom{5}{k}(0.1)^k(1 - 0.1)^{5-k}.$$

Following the formula of expected value (1), we have

$$E(X) = \sum_{\text{all } k} kP(X = k) = \sum_{k=0}^{5} k \times \left[\binom{5}{k}(0.1)^k(1 - 0.1)^{5-k}\right] = 0.5.$$

We may calculate this by hand, when the number of values X can take is small. But what if we have a bag of 1000 M&M's? We also provide you the R code.

R Code

```r
# possible values that X can take
values = 0:5

# probability for each value
probs = dbinom(values, size = 5, prob = 0.1)

# expected value
expected = sum(values * probs)
expected
```

```
## [1] 0.5
```

0.5 happans to be the product of the total number $n = 5$ and the probability of success $p = 0.1$. This is not a coincident, as in the table we provided in last week's review, the mean of a Binomial distribution of np. Here we just have shown you a particular case.

We will use expected value to calculate the expected loss associated with different hypothesis after seeing the data in this week's Decision Making Module.

Joint (Probability) Distribution (new)

When we have more than 1 random variable, say X and Y. The two may not be independent. Instead they may correlate to each other in some sense. If we want to consider them together as a pair, we need to introduce a "joint" probability distribution that captures the behavior of both X and Y.

For example, when considering to assign prior distribution to the parameters μ and σ of the Normal distribution, if we do not know σ, the Normal-Normal conjugate family cannot be used. Then we would need to consider a joint distribution of both μ and σ.

In the discrete case, the joint distribution of discrete random variables X and Y, is given by

$$p(k, l) = P(X = k \text{ and } Y = l).$$

Recall that, in general $P(X = k \text{ and } Y = l) \neq P(X = k) \times P(Y = l)$. ($X$ and Y may not be independent!) When one or more of the random variables are continuous, the joint distribution will be more complicated and cannot be easily calculated. We usually denote it as $p(x, y)$.

Marginal Distribution (new)

While we often cannot obtain the joint distribution $p(x, y)$ from the distribution $p(x)$ and $p(y)$, the other way is possible. That is, we can recover the distributions of X and Y: $p(x)$ and $p(y)$ from the joint distribution. And they are called the marginal distributions. Marginal distribution gives the probabilities of various values of the variables in the subset **without** reference to the values of the other variables. This contrasts with conditional probability, where we get the probability of the variables based on the values of other variables.

In the discrete case, we have the marginal distributions of X and Y to be

$$
\begin{aligned}
p(k) = P(X = k) &= \sum_{\text{all } y \text{ values}} P(X = k \text{ and } Y = l) = \sum_{\text{all } l} p(k, l) \\
p(l) = P(Y = l) &= \sum_{\text{all } x \text{ values}} P(X = k \text{ and } Y = l) = \sum_{\text{all } k} p(k, l)
\end{aligned}
$$

The continuous counterparts are given as

$$
\begin{aligned}
p(x) &= \int_{-\infty}^{\infty} p(x, y)\, dy \\
p(y) &= \int_{-\infty}^{\infty} p(x, y)\, dx
\end{aligned}
$$

While we may use the `integrate` function in R to help us to perform integrals, in this course, **we do not require any integration**. We will provide you the result directly.

Student's t-Distribution

In Course 2 **Inferential Statistics**, we used Student's t-distribution for inference of means. We say, only when we know the population standard deviation σ and the data pass all the assumptions, that we can use the z-score. Otherwise, we need the t-score to help us in hypothesis testings.

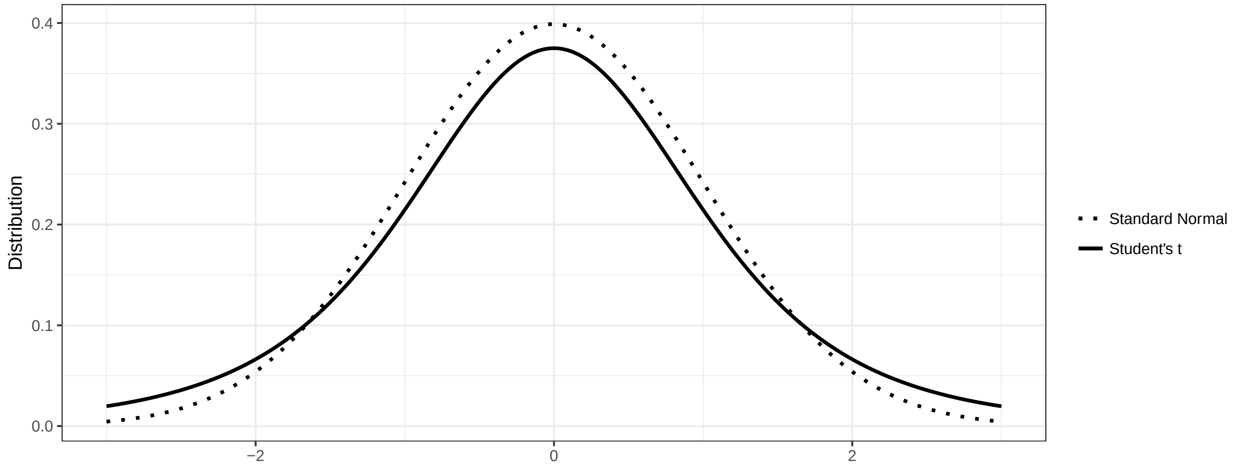
Here, we also provide you the formula of the Student's t-distribution, and show you how the t-score *standardize* the t-distribution. You do not need to be proficient with these details, because we have R functions to help us. However, they will be helpful if you are interested to know more about Student's t-distribution and how it relates to other distributions that we are more familiar with.

The standardized Student's t-distribution is defined to be

$$
p(t) = \frac{1}{\sqrt{\pi\nu}} \frac{\Gamma(\frac{\nu+1}{2})}{\Gamma(\frac{\nu}{2})} \left(1 + \frac{t^2}{\nu}\right)^{-\frac{\nu+1}{2}} .
$$

The $\Gamma(\cdot)$ is the Gamma function that we have seen in Week 2. This Student's t-distribution is centered at 0 (the location parameter), with a scale parameter equal to 1, and degree of freedom parameter ν.

Compared to the Normal distribution, the curve of the Student's t-distribution has slightly heavier tails, and therefore, it is a little "shorter" in the middle.



t-score

In the Course **Inferential Statistics**, when we conducted hypothesis testing or inference on the mean parameter, without knowing the population standard deviation σ, or with not enough data to meet the Normal assumption, we would use the t-score, instead of the z-score. Similar to the z-score, the t-score is defined to be some sort of a re-normalization:

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}}.$$

Here $\bar{x}$ is the sample mean of the random variable X, s is the sample deviation, and n is the sample size, and μ is the point estimate that we would like to infer about the population mean.

Student's t-distribution can be generalized into a distribution with 3 parameters: locatoin parameter m, scale parameter σ, and the degree of freedom parameter ν. The non-standardized Student's t-distribution is given by

$$p(x) = \frac{1}{\sigma\sqrt{\pi\nu}} \frac{\Gamma(\frac{\nu+1}{2})}{\Gamma(\frac{\nu}{2})} \left(1 + \frac{1}{\nu}\left(\frac{x-m}{\sigma}\right)^2\right)^{-\frac{\nu+1}{2}}. \tag{3}$$

We denote this t-distribution to be $\mathtt{t}(\nu, m, \sigma^2)$. Hence, the standardized t-distribution is represented as $\mathtt{t}(\nu, 0, 1)$. Using the transformation $y = \dfrac{x-m}{\sigma}$, we can recover the standardized t-distribution. That is, random variable $X \sim \mathtt{t}(\nu, m, \sigma^2)$, is equivalent to, random variable $Y = \dfrac{X-m}{\sigma} \sim \mathtt{t}(\nu, 0, 1)$.

The *Inferential Statistics* course tells us, the t-score, or t-statistics follows a standardized t-distribution $\mathtt{t}(n-1, 0, 1)$. Then the mean variable follows a non-standardized t-distribution $\mathtt{t}(n-1, \bar{x}, \frac{s^2}{n})$. The specialty of this case is, the degree of freedom is related to the scale parameter (both are involved with n). But it does not have to be the case for a non-standardized t-distribution.

Notice the σ in (3) is not necessarily the standard deviation of random variable X if X has distribution function $p(x)$ defined in (3). We only adopt this notation for the convention of σ used in the t-score.

Cauchy distribution (new)

A special t-distribution we will also use in Week 3 is called the Cauchy distribution. Cauchy distribution is the t-distribution when the degree of freedom $\nu = 1$

$$\mathsf{C}(m, \sigma^2) = \mathsf{t}(1, m, \sigma^2)$$

The formula for Cauchy distribution is

$$p(t) = \frac{1}{\sigma\pi}\left(1+\left(\frac{x-m}{\sigma}\right)\right)^{-1} = \frac{1}{\sigma\pi}\frac{1}{1+\left(\frac{x-m}{\sigma}\right)}.$$

R Code

Since t-distribution involves special function $\Gamma(\cdot)$ and complicated arithmetics, we will not expect you to know how to do the calculation by hands. R provides us the function `dt` and `dcauchy` for the **standardized** t-distribution and the general Cauchy distribution. To get the non-standardized t-distribution from the standardized one, we need to do the transformation

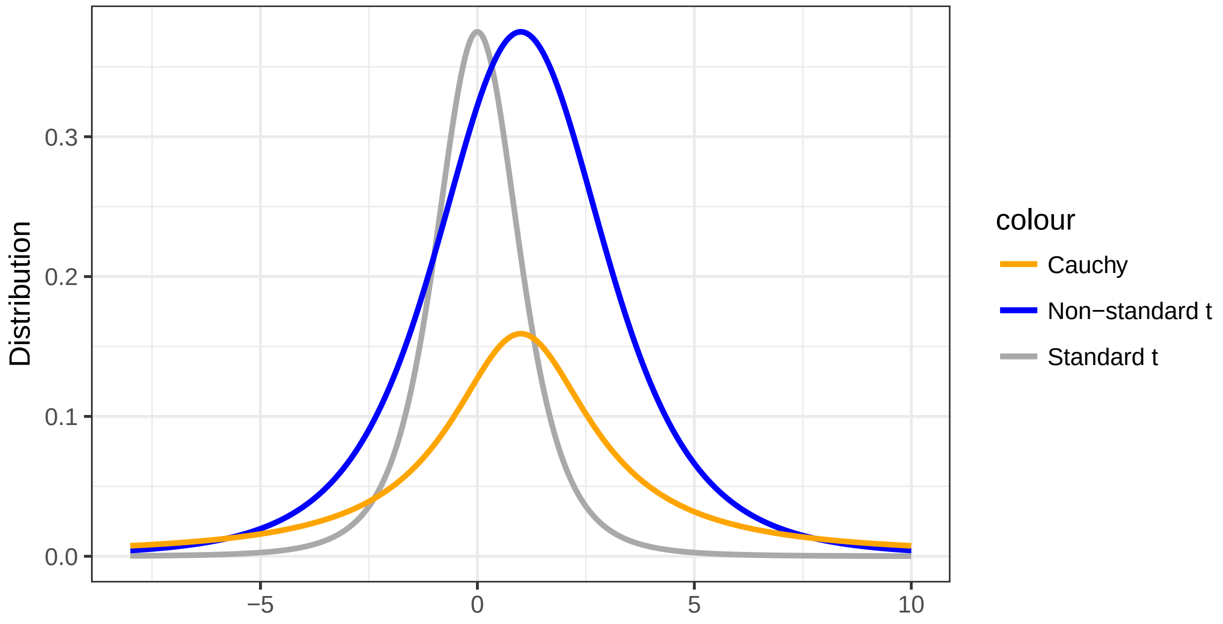
$$x = m + t \times \sigma, \qquad (\text{location} + t \times \text{scale}).$$

```
x = seq(-8, 10, by = 0.005)

# standardized t-distribution with degree of freedom 4
y_t_standard = dt(x, df = 4)

# non-standardized t-distribution with location 1, scale 2, degree of freedom 4
y_t_general = dt((x - 1) / 2, df = 4)

# Cauchy distribution with location 1, scale 2
y_cauchy = dcauchy(x, location = 1, scale = 2)
```



Hypothesis Test for Means

In this week, we will discuss hypothesis testing for means under Bayesian framework. Before we start, we would like to review the frequentist counterpart, so that we can compare the Bayesian approach shown in the lecture videos.

p-Value

Suppose we have two competing hypotheses, under the frequentist framework, we call one the *null hypothesis*,

$$H_0: \text{"nothing happens"},$$

and the other one, the *alternative hypothesis*,

$$H_A: \text{"something happens"}.$$

The p-value is defined to be, **given the null hypothesis is true**, the probability of seeing something or more extreme cases happen. We can use conditional probability to translate this into

$$p-\text{value} = P(\text{something or more extreme cases happen} \mid H_0 \text{ is true}).$$

We also set a significance level α. We say, if p-value is larger than α, we fail to reject the null hypothesis. On the other hand, if p-value is smaller than α, we reject the null hypothesis and accept the alternative hypothesis. α also gives the rate of Type I error, that is, we reject the null hypothesis when the null-hypothesis is true (known as a "false negative" finding).

However, p-value **does not provide the probability that H_0 is true**. Because of this fact, we **cannot** interpret a 95% confidence interval to be the probability of a parameter falling within this interval is 95%.

Hypothesis Test for Mean from One Sample

Suppose we want to infer the true value of a mean from one sample. We have the sample mean $\bar{x}$ and the sample standard deviation s. The sample size is n. We can set up the hypothesis test

$$H_0: \mu = \mu_0, \qquad H_A: \mu \neq \mu_0.$$

To get the p-value, we calculate the t-score

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}},$$

and look up the t-table for the row with degrees of freedom $n-1$. Since this is a two-sided test, we finally need to multiply the result by 2.

R Code

Under R, we can use the `t.test` function from the `stats` package to perform this test. Here `x` is the vector containing all the observations of the variable X of interest, `mu_0` is the inferred value μ_0. And we choose `alternative = "two.sided"` for a two-sided test.

```r
t.test(x, alternative = "two.sided", mu = mu_0, conf.level = 0.95)
```

The function we provided you in the `statsr` package in the second course *Inferential Statistics* is based on the `t.test` function.

Hypothesis Test for Two Paired Means

For two paired means test, we focus on the difference between the two variables. Let D be the difference between the two variables, and $\bar{D}$ be sample mean of this difference. s is the sample standard deviation of the difference, and n is the sample size. We set up the two hypotheses

$$H_0 : \mu_1 = \mu_2, \qquad H_A : \mu_1 \neq \mu_2,$$

which is equivalent to,

$$H_0 : D = \mu_1 - \mu_2 = 0, \qquad H_A : D \neq 0.$$

We calculate the t-score

$$t = \frac{\bar{D} - 0}{s/\sqrt{n}} = \frac{\bar{D}}{s/\sqrt{n}},$$

and look up the t-table for the row with degrees of freedom $n-1$. Since this is a two-sided test, we finally need to multiply the result by 2.

R Code

We again use `t.test` function from the `stats` package. Here `diff` is the difference between the two vectors `x1` and `x2`.

```r
diff = x1 - x2
t.test(diff, alternative = "two.sided", mu = 0, conf.level = 0.95)
```

Hypothesis Test for Two Independent Means

The hypothesis test for two independent means is still formulated as

$$H_0 : \mu_1 = \mu_2, \qquad H_A : \mu_1 \neq \mu_2.$$

However, it is a little more complicated than the two paired means, since we may not assume both variables have the same variance. In the course *Inferential Statistics*, we provided you one method to perform the t-test, that is, we still consider the difference $D = \mu_1 - \mu_2$, with a modified variance. Here we are not going to review the entire process. We just provide the R code for reference

R Code

Suppose we are interested in comparing vectors `x1` and `x2`. Then the test can be done by

```r
t.test(x1, x2, alternative = "two.sided", mu = 0, conf.level = 0.95)
```

Problem with p-Value

The logic under the frequentist framework is that, suppose H_0 is true, then p-value must be smaller than the significance level. Now if p-value is larger than the significant level, then H_0 must be false. However, we have never justified the given assumption, that

> suppose H_0 is true, then the p-value must be smaller than some given significance level.

Suppose we wanted to know whether people preferred Facebook to Twitter, and set the significance level to be $\alpha = 0.05$. We randomly selected users to survey, and we did this 20 times, each with the same sample size. Each time we calculated the corresponding p-value. The probability of getting at least 1 survey giving a significant p-value can be formulated as a Binomial process, with $n = 20$ "trials", "probability of success" $p = 0.05$. We can calculate by using 1 subtract the probability of the complementary event

$$P(k \geq 1) = 1 - P(k = 0) = 1 - \binom{20}{0} 0.05^0 (1 - 0.05)^{20} \approx 0.64.$$

This is a very high probability. What if we did just one survey, and we happened to run into this exactly one survey which gave us a significant p-value? In this case, the probability for p-value to be significant is 1, since it did happen. But can we declare that people must prefer Facebook to Twitter based on just this one significant p-value?

You may say, let us decrease the significance level α. However, α cannot be 0 (the ideal case), or we cannot even perform a hypothesis test.

For hypothesis testing, the main question we want to answer is, given the data, what the probability for a hypothesis to be true, which is

$$P(H_0 \text{ is true} \mid \text{observed data}), \tag{4}$$

but not the other way around (the p-value definition). Under the Bayesian framework, we see that (4) is exactly the posterior probability of a hypothesis H_0 after seeing the data. In this week, we will address the question, how to compare two competing hypotheses using their posterior probabilities, or the ratio of their posterior probabilities.