Master of Data Science Online Programme
Course: Basic Statistics
SGA #2: What makes you happy?

Student: Andrei Batyrov (Fall 2022)
arbatyrov@edu.hse.ru

Higher School of Economics (HSE)
National Research University
Faculty of Computer Science

May 29, 2023

Contents

List of Figures

Problem 1

Assume you perform a study to detect how using social networks affects people's happiness level. You have 20 volunteers. Your study is planned as follows. All participants are known to be active users of social networks. First you ask every participant to fill in special questionary that allows you to estimate their happiness level. After that, all participants will avoid using of social networks for one week. After this week, they complete similar questionary to detect their new level of happiness. Then, for each participant, their new happiness level is compared the initial one. Assume that for each participant their happiness level is changed: either decreased or increased. Let X be the random variable that models the number of participants for who increased their happiness level. Let $X_{obs} = 16$, i.e. 16 out of 20 participants become happier, and it's the only data on which you can make a decision. Your significance level is 5%.

1. You should state the null hypothesis and the alternative hypothesis of your research and explain your choices.
2. You should state how X is distributed provided that null hypothesis holds.
3. Would you claim that people become happier when they avoid using social networks based on this data?

Also keep in mind to provide any necessary calculations (p-values, etc.)

Solution

Since, as per our experiment, each participant may become either happier (happiness level increased) or less happy (happiness level decreases), one single participant's response can be modeled as the Bernoulli random variable with sample space having two elementary outcomes $\Omega = \{ I, D \}$, where I is the "Increased happiness" outcome and D is the "Decreased happiness" outcome. Let's assume that the volunteers do not know each other and respond independently. Then the sum of all their responses will have Binomial distribution, i.e. $X \sim Bin(n, p)$, where n is the number of volunteers, equal to 20 in our case, and p is the probability of "success" outcome, let in our case p be the probability of the "Increased happiness" outcome. A possible realization of X with $X_{obs} = 16$ might be as follows: $x = (D, I, I, I, I, I, I, D, I, I, D, I, I, I, I, I, D, I, I, I)$, i.e. the first participant is less happy after the experiment, the second one is happier after the experiment, and so on.

Now since we do not know much about the participants and assuming that they are independent and active users of social networks, it is reasonable to assume that the probabilities of both outcomes are equal, i.e. $p = 1/2$. That is, half of them may become happier after the experiment, while the other half may become less happy. This is our null hypothesis. Since we want to understand if the experiment was successful, our alternative hypothesis is that the number of happier volunteers has increased, i.e. $p > 1/2$:

$\mathcal{H}_0 : p = 1/2$, happier and less happy outcomes are equally probable,

$\mathcal{H}_1 : p > 1/2$, happier outcomes are more probable.

The theoretical distribution of $X \sim Bin(n = 20, p = 0.5)$, provided that $\mathcal{H}_0$ holds. Given X_{obs}, we need to decide, if we should reject the null hypothesis, that is we want to evaluate how unusual it is to get X_{obs}, provided that the null hypothesis is true (I and D outcomes are equally probable). To make this decision, we need to either pass or fail the Binomial test. First we need to find the so-called p-value for that X_{obs}. As per [3]

$$p\text{-value}(X_{obs}) = P(X \geq X_{obs} | \mathcal{H}_0). \tag{1}$$

That is p-value is the probability of obtaining a value of X at least equal to X_{obs} or greater, provided that the null hypothesis holds. Let's find this probability. $P(X \geq X_{obs}) = \sum_{k=X_{obs}}^{n} P(X = k)$. As per [1], $P(X = k)$ is the probability mass function of the binomial distribution and is

$$P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}, \tag{2}$$

where p is the probability of "success" (happiness level increased), n is the total number of participants, k is the number of "happier" responses. Since $p = 0.5$, $1 - p = 1 - 0.5 = 0.5 = p$, (2) can

be re-written as

$$P(X=k)=\binom{n}{k}p^k(1-p)^{n-k}=\binom{n}{k}p^kp^{n-k}=\binom{n}{k}p^n=\frac{n!}{k!(n-k)!}p^n. \tag{3}$$

So, $P(X\geq X_{obs})=p^n n!\sum\limits_{k=X_{obs}}^{n}\frac{1}{k!(n-k)!}$. And finally,

$$p\text{-value}(X_{obs}=16)=\left(\frac{1}{2}\right)^{20}20!\left(\frac{1}{16!4!}+\frac{1}{17!3!}+\frac{1}{18!2!}+\frac{1}{19!1!}+\frac{1}{20!0!}\right)=\frac{1549}{262,144}\approx 0.6\times10^{-2}. \tag{4}$$

Second, we need to compare the result (4) with the significance level $\alpha=5\%=5\times10^{-2}$. As per [6], the significance level α is the probability of Type I error, that is the probability to falsely reject the null hypothesis while it is true. Type I error also means yielding a false positive result, i.e. reporting some effect which is not present, in fact. It is easily seen, that p-value $\approx 0.6\times10^{-2}<\alpha=5\times10^{-2}$, which means that we are well within the critical region of X and we control the probability of the Type I error not exceeding the significance level of 5%. It is indeed unusual to get $X_{obs}=16$, provided that $\mathcal{H}_0$ is true (I and D outcomes are equally probable). The result is statistically significant. Strictly speaking, we only reject $\mathcal{H}_0$ and this does *not* mean that we automatically should accept alternative hypothesis(es). In other words, we have only shown that we control the probability of Type I error, i.e. we do not falsely report some effect which is not present. However, since we used only one alternative hypothesis (one-tail test) for evaluating if $p>1/2$, we can reject the null hypothesis in favor of that alternative hypothesis $\mathcal{H}_1: p>1/2$ (happier outcomes are more probable). In this case, we may claim that people become happier when they avoid using social networks based on the given data.

Further Considerations

We have just shown that $X_{obs}=16$ is statistically significant result by passing the Binomial test. Let's confirm this result by passing another statistical test. For large samples, due to the Central Limit Theorem, we might use the Normal Distribution to model our experiment. However, since we have a rather small number of participants $n=20>1$, but not $n=20\gg1$, let's consider one-sample Student's t-test [4], for which our random variable X is assumed to be distributed according to the Student's t-distribution. This statistical test may be used for small samples with existing but unknown population variance. In this case the sample variance is used instead. Student's t-distribution uses three parameters: mean, standard deviation and degrees of freedom. The null and alternative hypotheses for our one-sample t-test are

$\mathcal{H}_0: \mu=\mu_0$, the sample mean obtained after the experiment equal to the population mean,

$\mathcal{H}_1: \mu>\mu_0$, the sample mean obtained after the experiment is greater than the population mean.

As per results obtained in previous assignments for the Probability Theory course, we know that for the Binomial distribution $\mathbb{E}X=np$ – this is the population mean μ_0. And the variance is $Var(x)=np(1-p)$ – this is our sample variance, and sample standard deviation is $std(x)=\sqrt{Var(x)}$. Degrees of freedom [5] is a parameter that depends on the sample size and is $\nu=n-1$. As with the Binomial test, we need to find the p-value and compare it with the significance level α. Recalling that as per [2] $P(X\leq X_{obs})$ is the value of cumulative distribution function at X_{obs}:

$$p\text{-value}(X_{obs})=P(X\geq X_{obs}|\mathcal{H}_0)=1-P(X<X_{obs})=1-P(X\leq X_{obs}-1)=1-CDF(X_{obs}-1), \tag{5}$$

where CDF is the cumulative distribution function of the t-distribution. We have to exclude the value X_{obs} from the probability for CDF, since it is already included in the probability for p-value. CDF for t-distribution cannot be expressed in elementary functions and is usually computed with numerical methods. Let's use the Python's library Scipy to find the value of $CDF(X_{obs}-1)$ for t-distribution with $\mu_0=np=20\times0.5=10$, $std(x)=\sqrt{Var(x)}=\sqrt{np(1-p)}=\sqrt{20\times0.5\times0.5}=\sqrt{5}$, and $\nu=n-1=20-1=19$. Finally,

$$p\text{-value}(X_{obs}=16)=1-CDF(16-1)\approx1-0.98123=0.01877\approx1.88\times10^{-2}. \tag{6}$$

Again, it is easily seen, that p-value $\approx 1.88\times10^{-2}<\alpha=5\times10^{-2}$, which means that we are well within the critical region of X and we control the probability of the Type I error not exceeding the significance level of 5%. It is indeed unusual to get $X_{obs}=16$, provided that $\mathcal{H}_0$ is true (the sample mean obtained after the experiment equal to the population mean). The result

is statistically significant. Again, as with the Binomial test, since we used only one alternative hypothesis (one-tail test) for evaluating if $\mu > \mu_0$, we can reject the null hypothesis in favor of that alternative hypothesis $\mathcal{H}_1 : \mu > \mu_0$ (the sample mean obtained after the experiment is greater than the population mean). In other words, we may claim that the increased number of happier people ($X_{obs} = 16$) as compared to the expected number of happier people $\mu_0 = 10$ after avoiding using social networks is statistically significant.

So, both tests have confirmed the statistical significance of the result of our experiment. Now let's take a look at the plots of probability mass function of the Binomial distribution and the probability distribution function of the t-distribution – see Figure 1. Visually we can confirm that t-distribution is a good approximation for Binomial distribution in our case. However, let's take a closer look at the right tail of both – see Figure 2. The p-value($X_{obs} = 16$) for the t-distribution is shown as the shaded area under the pdf curve. It is easily seen, that the t-distribution has heavier tails, particularly the right tail – the pdf curve lies above the pmf points. This means that, for example, by letting $\alpha = 10^{-2}$, the t-test would fail, since p-value $\approx 1.88 \times 10^{-2} > \alpha = 10^{-2}$. In this case, to pass the test we may want to have more unusual observations – shift X_{obs} to the right – to get lower p-value to satisfy the significance level. It is easy to check that by letting $X_{obs} = 17$, i.e. 16 happier volunteers plus one more, we get p-value($X_{obs} = 17$) $= 1 - CDF(17 - 1) \approx 1 - 0.992645 = 0.007355 \approx 0.74 \times 10^{-2} < \alpha = 10^{-2}$. This mean that t-test looks like a more robust test for confirming statistical significance, since due to the t-distribution having heavier tails, it requires more unusual observations to control the probability of Type I error (reporting some effect which is not present, in fact).

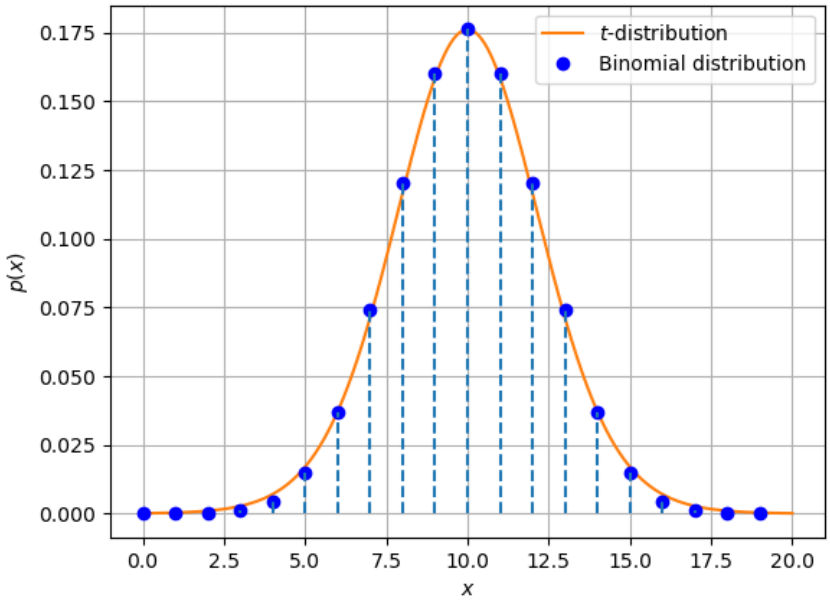

Figure 1: Plots of Binomial's pmf and t-distribution's pdf

Answer

1. The null hypothesis is "happier and less happy volunteers are equally probable". This seems reasonable, since according to the experiment, the responses may be modelled as the Bernoulli process with equally probable outcomes. Since we want to understand if the experiment was successful, i.e. happiness level has increased, the alternative hypothesis is "happier volunteers are more probable".

2. Our random variable X is distributed as $X \sim Bin(n = 20, p = 0.5)$, provided that null hypothesis holds.

3. We have run two tests: Binomial test and t-test, and both confirmed that people become *happier* when they avoid using social networks: p-value($X_{obs} = 16$) $\approx 0.6 \times 10^{-2}$ for the Binomial test and p-value($X_{obs} = 16$) $\approx 1.88 \times 10^{-2}$ for the t-test, both values are less than

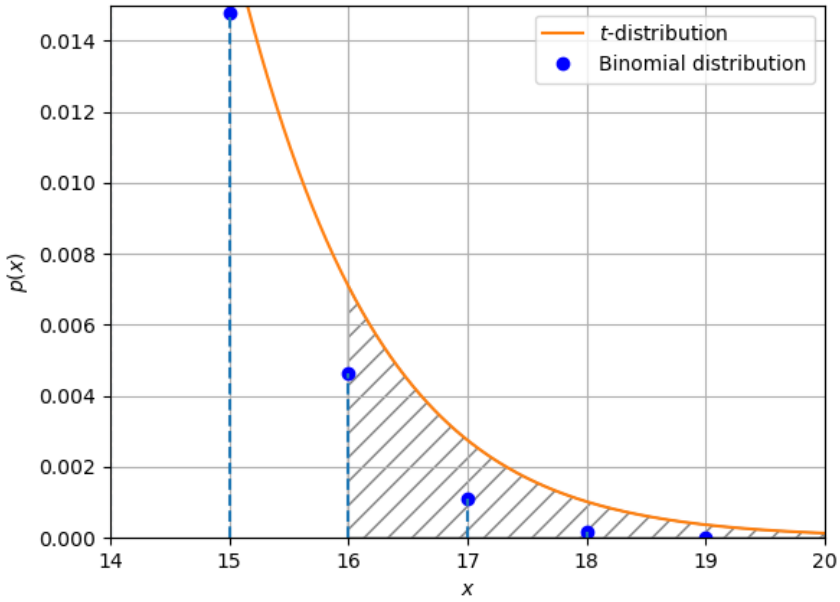

Figure 2: Distributions' right tails with shown p-value for t-distribution

the significance level $\alpha = 5 \times 10^{-2}$. Also, the t-test looks to be more robust, since it requires more unusual observations to control the probability of Type I error (reporting some effect which is not present, in fact).

References

[1] Ilya Schurov. *Binomial distribution*. Faculty of Computer Science, Higher School of Economics. URL: https://smartedu.hse.ru/mod/page/0/756754.

[2] Ilya Schurov. *Cumulative distribution function (CDF)*. Faculty of Computer Science, Higher School of Economics. URL: https://smartedu.hse.ru/mod/page/0/756817.

[3] Ilya Schurov. *Introducing p-value*. Faculty of Computer Science, Higher School of Economics. URL: https://smartedu.hse.ru/mod/page/0/798133.

[4] Ilya Schurov. *One-sample Student's t-test*. Faculty of Computer Science, Higher School of Economics. URL: https://smartedu.hse.ru/mod/page/0/798143.

[5] Ilya Schurov. *T-distribution*. Faculty of Computer Science, Higher School of Economics. URL: https://smartedu.hse.ru/mod/page/0/798145.

[6] Ilya Schurov. *Type I and type II errors*. Faculty of Computer Science, Higher School of Economics. URL: https://smartedu.hse.ru/mod/page/0/798130.