Biostats Lecture 5: Statistical Hypothesis Testing

Public Health 783

Ralph Trane
University of Wisconsin–Madison

Fall 2019

WISCONSIN
UNIVERSITY OF WISCONSIN–MADISON

Recap

Random variables

Distributions

Estimators

Estimators are random variables! (for example, the average is a random variable)

Statistical Hypothesis Testing

Scenario:

We've been playing a simple game. Everytime you roll a six, I pay you a dollar. Everytime I roll a six, you pay me a dollar.

I've had crazy good luck, and by the end of the day won a lot of money from you.

You accuse me of cheating, and demand to test the dice I've been using!

I agree to let you test them, but ONLY if you do it in a sound, statistical manner. How to go about that?

You decide to roll the dice 3 times each, for a total of 27 rolls. You assume they'll all behave the same, so the probability of rolling a six is the same for all three dice.

Statistical Hypothesis Testing

Setup:

Let $X_1, X_2, \ldots, X_{27}$ be the outcomes of the thirty "trials". Each trial consists of rolling a die, and check if it's a six or not. If it's a six, we'll call it a success, if not we'll call it a failure. I.e. $X_i \sim \text{Bernoulli}(\pi)$.

IF the dice are fair, $P(X_i = 1) = 1/6$ for all $i = 1, 2, \ldots, 27$.

IF the dice are fair, we would expect to roll a 6 close to $\frac{1}{6} \cdot 27 = 6$ times, i.e. about 5 of the X's should be 1's.

What would cause you to reject the idea that the dice are fair?

If we see way more than 6 sixes. What would be "way more"? 7? 8? 17?

Statistical Hypothesis Testing

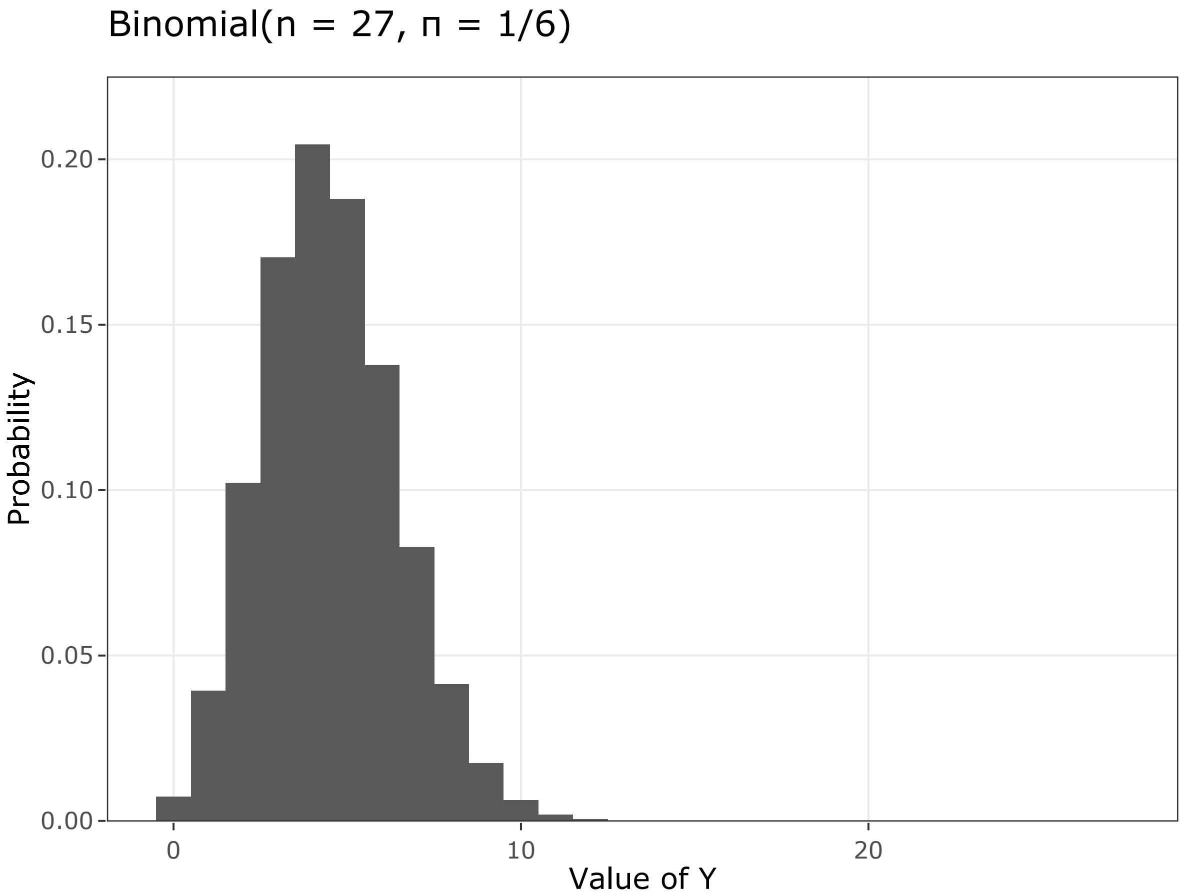
In terms of probabilities: what is the *probability* of observing at least 10 sixes *IF* the $P(X_i = 1) = \frac{1}{6}$?

- if the probability is small, 10 is a lot of sixes
- if the probability is large, 10 is a reasonable number of sixes

First, introduce the random variable $Y =$ number of sixes $= X_1 + X_2 + \ldots X_{27}$. The probability of observing more than 10 sixes is $P(Y \geq 10)$. To find this, we need the distribution of Y, which is $\text{Binomial}(27, \pi)$, where π is the probability of rolling a six.

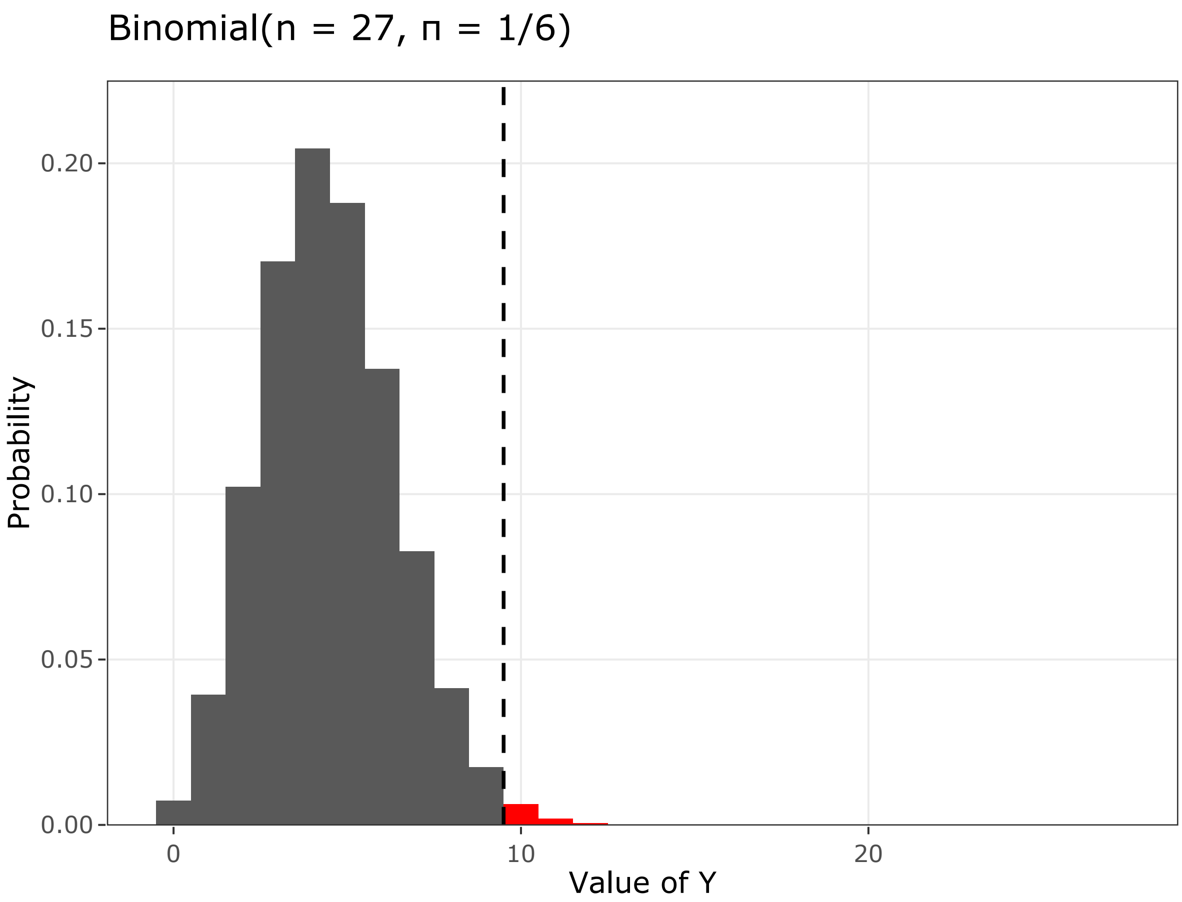
Statistical Hypothesis Testing

IF the dice are fair, $\pi = \frac{1}{6}$. So *IF* the dice are fair, the distribution of Y looks like this:



Statistical Hypothesis Testing

The probability we want to find is the red area below. We will do this in SAS in just a second. The result is 0.00261.



Statistical Hypothesis Testing

This means that *IF* the true probability of rolling a six with these dice is indeed $\frac{1}{6}$, the probability of rolling 10 or more sixes is 0.00261. This probability is called the *p-value* for the test $H_0 : \pi = 1/6$ when testing against $H_A : \pi > 1/6$.

Is this small enough to convince you that the true probability is *NOT* $\frac{1}{6}$?

Statistical Hypothesis Testing

In SAS.

Different (but the same) approach

Instead of looking at Y (number of sixes), look at the proportion of sixes. I.e.

$$\hat{p} = \frac{Y}{n} = \frac{1}{27}(X_1 + \ldots + X_{27}) = \frac{1}{27}\sum_{i=1}^{27} X_i$$

This is an average, so CLT tells us it's normally distributed around the true value of π, and variance $\frac{\mathrm{Var}(X_i)}{n} = \frac{\pi(1-\pi)}{27}$.

We reject the idea that the true value of π is $\frac{1}{6}$ when $\hat{p}$ is "far from" $\frac{1}{6}$.

I.e. when $\hat{p} - \frac{1}{6}$ is large. When this is large, so is $Z = \frac{\hat{p} - 1/6}{\mathrm{SD}(\hat{p})}$.

If we pretend $\pi = \frac{1}{6}$, then $Z \sim N(0, 1)$!

If we pretend $\pi = \frac{1}{6}$, then $\mathrm{SD}(\hat{p}) = \sqrt{\mathrm{Var}(\hat{p})} = \sqrt{\frac{1/6 \cdot (1 - 1/6)}{27}}$, and so we can actually calculate z_{obs}, the observed value of Z.

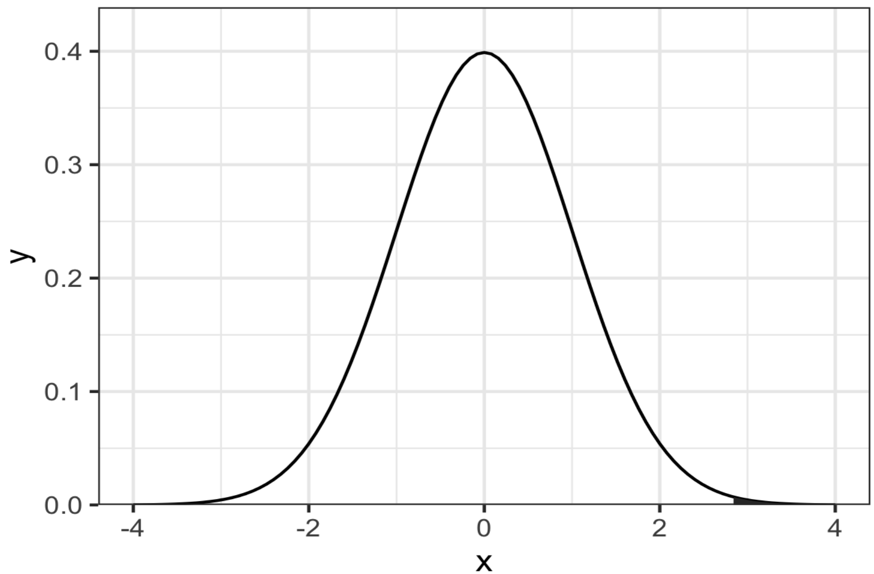
Different (but the same) approach

So,

$$\begin{aligned}\text{p-value} &= P(\hat{p} > p_{obs}) \\ &= P\left(\frac{\hat{p} - 1/6}{\text{SD}(\hat{p})} > \frac{p_{obs} - 1/6}{\sqrt{1/6 \cdot (1 - 1/6)/27}}\right) \qquad = P(Z > 2.84019)\end{aligned}$$

Statistical Hypothesis Testing

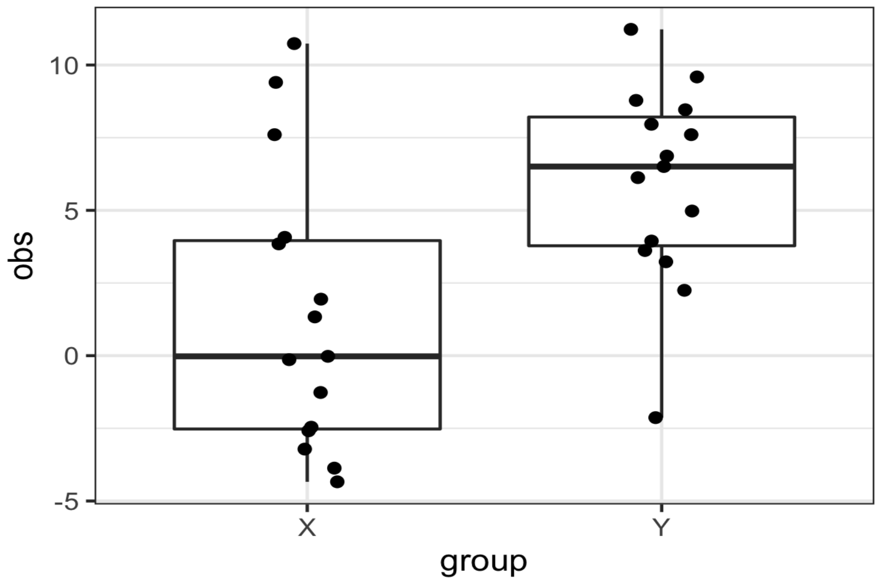
So we found our p-value as the area depicted below. Turns out, this is 0.00225.



This strategy can be used every time we the quantity of interest in follows a (at least approximately) normal distribution. Subtract the mean and divide by the standard deviation to get the standard normal. Then find the probability.

Group Comparison

Two groups, 15 observations in each group. Want to test $H_0 : \mu_X = \mu_Y$ against $H_A : \mu_X \neq \mu_Y$.



Statistical Hypothesis Testing

Natural to look at $\bar{X} - \bar{Y}$.

If n is "large enough", both of these quantities are normally distributed. From last week, normal minus normal is normal. So $\bar{X} - \bar{Y}$ is normal!

Next question: exactly what normal would it be *IF* we pretend the null hypothesis is true?

Mean: $E(\bar{X} - \bar{Y}) = E(\bar{X}) - E(\bar{Y}) = \mu_X - \mu_Y = 0$

Variance (assuming independence): $\text{Var}(\bar{X} - \bar{Y}) = \text{Var}(\bar{X}) + \text{Var}(\bar{Y}) = \frac{\sigma_X^2}{n_X} + \frac{\sigma_Y^2}{n_Y}$.

So $Z = \frac{\bar{X}-\bar{Y}}{\text{SD}(\bar{X}-\bar{Y})} = \frac{\bar{X}-\bar{Y}}{\sqrt{\text{Var}(\bar{X}-\bar{Y})}} \sim N(0,1)$.

What's the problem here? We don't know σ_X^2 or σ_Y^2!

Statistical Hypothesis Testing

Luckily, we can simply *estimate* both, and plug them in.

When we do this, the result no longer follows a standard normal distribution… BUT it follows a particular t-distribution. How to calculate the exact degrees of freedom is a bit tricky, but it can be done.

For the patient type:

The degrees of freedom is

$$\mathrm{df} = n_X + n_Y - 2$$

if we are willing to assume $\sigma_X^2 = \sigma_Y^2$. In this case we would use $s_X^2 = s_Y^2 = \frac{1}{n_X+n_Y-2}\left(\sum_{i=1}^{n_X}(x_i - \bar{x})^2 + \sum_{i=1}^{n_Y}(y_i - \bar{y})^2\right)$.

Statistical Hypothesis Testing

If you're not willing to assume $\sigma_X^2 = \sigma_Y^2$, the degrees of freedom is

$$\mathrm{df} = \frac{\left[\frac{s_X^2}{n_X} + \frac{s_Y^2}{n_Y}\right]^2}{\frac{(s_X^2/n_X)^2}{n_X-1} + \frac{(s_Y^2/n_Y)^2}{n_Y-1}},$$

and we would then use the regular sample variances as estimates:

$$s_X^2 = \frac{1}{n_X - 1}\sum_{i=1}^{n_X}(x_i - \bar{x})^2$$

$$s_Y^2 = \frac{1}{n_Y - 1}\sum_{i=1}^{n_Y}(y_i - \bar{y})^2$$

For the not so patient type:

Don't worry about it, you'll use software to calculate this.

HOWEVER, you will need to know WHEN you can assume equal variance.