7

Math in Economics Series

Reviewing Statistics

This tutorial covers:

- a random variable and its probability distribution
- mean, variance and standard deviation
- independence between two random variables
- correlation and covariance
- normal, chi-Squared, t, and F distributions
- hypothesis testing and confidence intervals

1 What Is a Random Variable?

A **random variable** is a variable with values that depend on the outcome of a random process. That is, a random variable measures a random outcome. For example, if we flip a coin 5 times, we know that the number of heads will be a whole number in the range from 0 to 5. Before we flip the coin the first time, however, we don't know what the outcome of any flip will be. In this example of coin-flipping, the number of heads in 5 flips is a random variable.

The number of heads in 5 flips is a **discrete random variable** because it takes on a finite number of values: 0, 1, 2, 3, 4, 5. The **probability distribution** of a discrete random variable is a list of the values of all possible outcomes and the corresponding probability of each outcome.

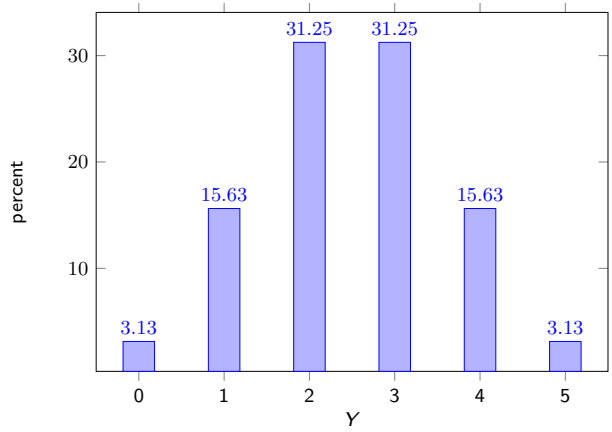
Suppose we flip a fair coin once. The outcome is either head or tail. Let's denote the number of heads (0 or 1) as variable X. $X = 1$ if it's head, and $X = 0$ if it's tail. The probability that the coin flip is a head is one half; formally, $P(X = 1) = 1/2$. And the probability that the outcome is a tail is also one half; $P(X = 0) = 1/2$. (We read "$P(X = 1) = 1/2$" as "the probability that X equals 1 is one half.") The probability distribution of the discrete random variable X is

$$X = \begin{cases} 0 & \text{with probability } 1/2 \\ 1 & \text{with probability } 1/2 \end{cases}$$

Since there are only two possible outcomes of a coin flip, the two probabilities sum to one.

Let's have Y denote the number of heads that occur in 5 coin flips, so Y is the random variable in our original problem. This figure displays the probability distribution of the random variable

Y in this case.



Each probability lies between 0 and 1 (or 100%), and the probabilities sum to 1 (or 100%). Also notice that this distribution is symmetric around $Y = 2.5$ heads.

You might be curious where the probabilities come from. The formula doesn't concern us here, but this hint might satisfy your curiosity: for 5 heads in row, we have $\left(\frac{1}{2}\right)^5 = \frac{1}{32} = 3.13\%$.

2 Calculating Means and Variances

The **mean** or expected value of a discrete random variable is the average of all the possible values of the outcome, weighted by the corresponding probabilities. With a single coin flip, the mean of our random variable X is

$$\mu_X = E(X) = \left(0 \times \frac{1}{2}\right) + \left(1 \times \frac{1}{2}\right) = \frac{1}{2}.$$

On average, how far are the values of the random variable away from its mean value? The answer to this question is a measure of **dispersion**. In particular, if all the values of the random variable are quite close to the mean, then the distribution has little dispersion. A random variable's variance is a measure of dispersion that's based on average distance from the mean.

The **variance** of a discrete random variable is the expected value of the square of the difference between the variable and its mean. The variance of the random variable X is

$$\begin{aligned}\sigma_X^2 = VAR(X) &= \left[\left(0 - \frac{1}{2}\right)^2 \times \frac{1}{2}\right] + \left[\left(1 - \frac{1}{2}\right)^2 \times \frac{1}{2}\right] \\ &= \frac{1}{8} + \frac{1}{8} = \frac{1}{4}.\end{aligned}$$

The **standard deviation** of X is the square root of the variance. In fact, it's the square root of the weighted sum of distances, where each distance is $(X - \mu_X)^2$. So the standard deviation is the average distance of the values of the random variable from its mean. In the case of a single coin flip, the standard deviation is $\sigma_X = \frac{1}{2}$.

Practice Question

You are using your laptop to do your homework. Let C indicate whether your laptop crashes during an hour. The probability distribution of C is

$$C = \begin{cases} 0 & \text{with probability } 0.90 \\ 1 & \text{with probability } 0.10 \end{cases}$$

What are the mean, variance, and standard deviation of C?

3 Independence and Correlation

Two random variables are **independent** if knowing the outcome of one variable doesn't affect the probability distribution of the other, and vice versa. For instance, the coin-flip variable X and the computer-crash variable C are surely independent; that is, the outcome of flipping a coin once doesn't affect whether your laptop crashes this hour. Also, if you flip a coin repeatedly, the outcome of any flip is independent of the outcomes of the other flips. For instance, getting a head on the first flip doesn't change the laws of physics to make a head less (or more) likely in any subsequent flip; the probability remains $\frac{1}{2}$.

To measure how two random variables Y and Z are related, consider their **covariance**.

$$\sigma_{YZ} = COV(Y, Z) = E[(Y - \mu_Y)(Z - \mu_Z)]$$

If Y tends to be large when Z is large and small when Z is small, then the products $(Y - \mu_Y)(Z - \mu_Z)$ will tend to be positive, and $\sigma_{YZ} > 0$. If, however, Y tends to be small when Z is large and large when Z is small, then the products $(Y - \mu_Y)(Z - \mu_Z)$ will tend to be negative, and $\sigma_{YZ} < 0$.

Example. The unemployment rate tends to be high when economic growth slows and turns negative. In annual data for the United States since 1948, the covariance of the unemployment rate and the growth rate of real GDP is -1.37.

The **correlation coefficient** is a unit-free measure of association between two variables; values of the correlation coefficient range from -1 to 1. For the two random variables Y and Z, the correlation coefficient is

$$CORR(Y, Z) = \frac{\sigma_{YZ}}{\sigma_Y \sigma_Z}.$$

If we change the units of Y by multiplying values of that random variable by a constant b, the

correlation coefficient doesn't change: the numerator and denominator scale up by the same factor b, so the ratio isn't affected. This unit-free property of the correlation coefficient is an advantage over the covariance, which doesn't have the unit-free property.

If Y and Z are independent, then $CORR(Y, Z)$ and $COV(Y, Z)$ equal zero. Zero correlation, however, doesn't imply independence. If Y and Z are related nonlinearly, then they aren't independent, but their correlation coefficient could equal zero.

Practice Question

In the annual data on the unemployment rate and the real growth rate of GDP since 1948, the covariance of the unemployment rate u and the real growth rate g is -1.37, the standard deviation of u is 1.60, and the standard deviation of g is 2.29. What is the correlation coefficient for u and g in these data?

4 Continuous Distributions: Normal, Chi-Squared, *t*, and *F*

Unlike a discrete random variable, a continuous random variable can take on an unlimited (i.e., infinite) number of possible values. The probability distribution of a discrete random variable lists each possible value of that variable and the corresponding probability. With a continuous random variable, the list of outcomes would be endless, so using a list to describe the distribution of a continuous variable isn't an option. Alternatively, we plot the random variable's **probability density function** (or **p.d.f.**).

Rather than specify the probability that the outcome of a continuous random variable is a particular number, we measure the probability that the outcome falls on an interval between two points. Suppose we're studying annual consumption of ice cream per person. There's no chance that a person's consumption of ice in a year is 407.39847362593038 ounces, but we can find the probability that personal ice cream consumption is between 405 and 410 ounces per year. We find the probability of ice cream consumption falling between 405 and 410 ounces by computing the area under the p.d.f. from 405 to 410. And Figure 2.5 shows that the probability of a normal random variable being greater than $\mu + 1.96\sigma$ is 2.5%.

Four continuous distributions play leading roles in econometrics, but the star of the econometrics show in the **normal distribution**. A normal random variable is a continuous random variable. It's values cover all the real numbers (i.e., from $-\infty$ to $+\infty$). Figure 2.5 (Stock and Watson 2015) displays the familiar bell shape of the normal probability density function.

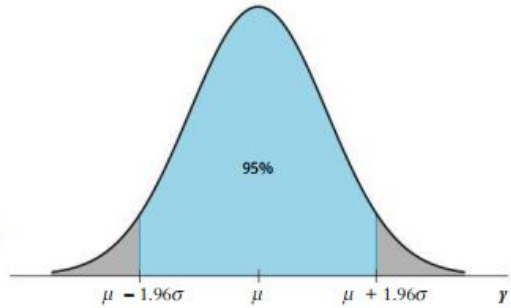
The probability density function of a normal random variable is symmetric with mean μ and variance σ^2: $N(\mu, \sigma^2)$. If the the mean of a normal random variable is 0 and its variance is 1, we have the **standard normal distribution** $N(0, 1)$.

If a variable has the standard normal distribution, the probability that the outcome falls below -1.96 is 2.5%, and the probability that the outcome exceeds 1.96 is 2.5%. That leaves 95% as the probability that Y lies between -1.96 and $+1.96$. Below, we'll see how these properties are

important for testing hypotheses.

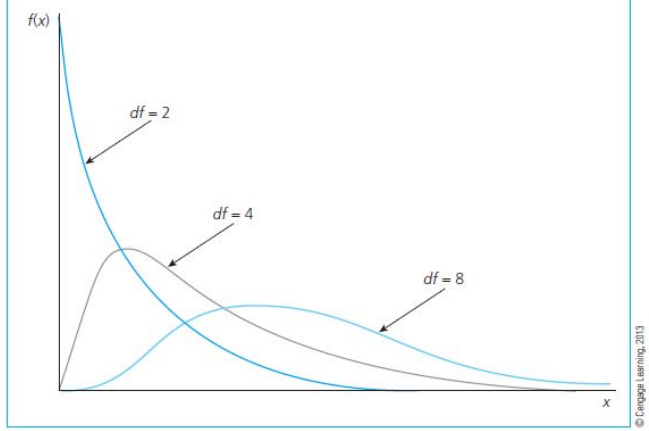
FIGURE 2.5 The Normal Probability Density

The normal probability density function with mean μ and variance σ^2 is a bell-shaped curve, centered at μ. The area under the normal p.d.f. between $\mu - 1.96\sigma$ and $\mu + 1.96\sigma$ is 0.95. The normal distribution is denoted $N(\mu, \sigma^2)$.



The **chi-squared distribution** is the distribution of the sum of squares of k independent standard-normal random variables, written as χ^2_k, and k is called the degrees of freedom. As a sum of squares, a chi-squared random variable can't be negative, so the distribution of a chi-squared random variable is defined on the domain from 0 to infinity. The shape of the chi-squared's p.d.f. depends on its degrees of freedom, which we see in Figure B.9 (Wooldridge 2013). And the chi-squared distribution is clearly not symmetric.

FIGURE B.9 The chi-square distribution with various degrees of freedom.



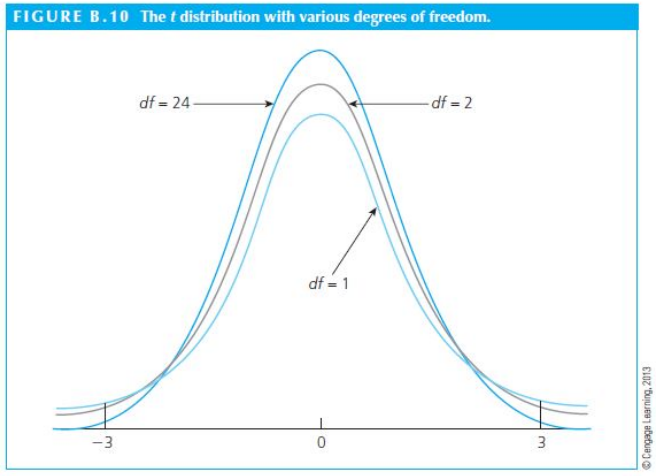
Suppose A, B, C, and D are three standard-normal random variables. Then $X = A+B+C+D$ is also a standard-normal random variable. And if A, B, C, and D are independent, $Y =$

$A^2 + B^2 + C^2 + D^2$ has a chi-squared distribution with $k = 4$ degrees of freedom. The $df = 4$ curve in Figure 8.9 is the probability density function of Y.

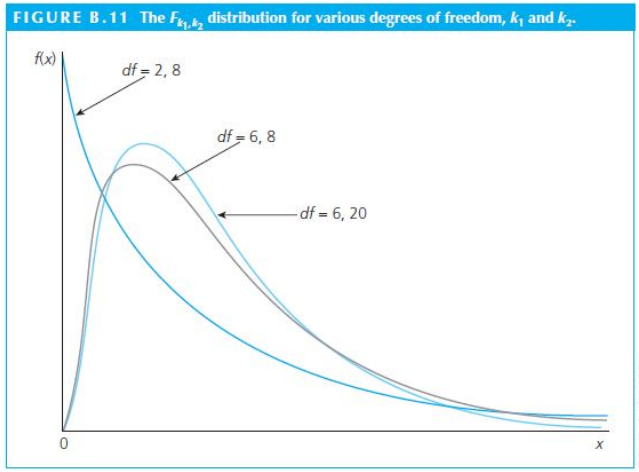
If a random variable Z is the ratio of a standard-normal random variable X to a chi-squared random variable Y, then the variable Z has a ***t* distribution** with n degrees of freedom.

$$Z = \frac{X}{Y}$$

The p.d.f. of the *t* distribution has the shape in Figure B.10 (Wooldridge 2013). It's symmetric, and a bit lower and wider than the standard normal p.d.f., so observations in the tails are more common. If n is large, the *t* distribution resembles the standard normal distribution.



An ***F* distribution** with (k_1, k_2) degrees of freedom is the ratio of one chi-squared random variable (divided by its degrees of freedom k_1) to another chi-squared random variable (divided by its degrees of freedom k_2), and the two random variables are independent. Figure B.11 (Wooldridge 2013) displays the shape of the F distribution for three pairs of degrees of freedom.



The F distribution is often used to test hypotheses in the context of multiple regressions.

5 Hypothesis Testing and Confidence Intervals

Researchers collect accounting data from thousands of McDonald's franchises, and, with great effort and deep understanding, they generate a measure of profit that reflects opportunity costs. In fact, for each franchise i in their sample of 5,000 McDonald's franchises, they measure economic profit π_i.

Is economic profit zero, as in a long-run competitive equilibrium? Or is profit positive or negative? In the population, profit π is a random variable with mean μ_π. We hope to determine from the sample of 5,000 franchises whether there's evidence to reject the hypothesis that $\mu_\pi = 0$.

Let's be specific about the test. The **null hypothesis** is

$$H_0 : \mu_\pi = 0.$$

The null hypothesis would be wrong if $\mu_\pi < 0$ or $\mu_\pi > 0$. So the two-sided alternative hypothesis is

$$H_1 : \mu_\pi \neq 0.$$

Testing this hypothesis would be easy enough if we knew the standard deviation of economic profit π—not in the sample but in the population. If we knew σ_π, we would create the normal statistic

$$z = \frac{\overline{\pi}}{\sigma_\pi / \sqrt{n}}$$

where $\overline{\pi}$ is the sample mean, and $n = 5000$ is the sample size (i.e., the number of observations). The Central Limit Theorem tells us that any sample mean is normally distributed in large sam-

ples, so $\overline{\pi}$ is normally distributed. Since the standard deviation of the mean is $\sigma_\pi/\sqrt{n}$, the ratio z has a standard normal distribution.

So if the standard deviation of profit were known, testing our hypothesis of zero economic profit would involve computing the sample mean $\overline{\pi}$, forming the ratio z, and checking whether the value of z is unusually small or large. If our standard for "unusual" is 5% (i.e., 2.5% in each tail), then we would reject the null hypothesis of zero profit if z in the sample is greater than 1.96 or less than -1.96.

The method is similar if we have to estimate the standard deviation of σ_π. In our sample, we compute the mean $\overline{\pi}$ and the standard deviation s_π. Our new test statistic is the t ratio.

$$t_n = \frac{\overline{\pi}}{s_\pi/\sqrt{n}}$$

Replacing the unknown standard deviation σ_π with the sample standard deviation s_π is easy enough. But the ratio t_n isn't distributed normally. t_n has a t distribution with n degrees of freedom. That means that the critical values for the hypothesis test come from the t distribution. Recall, however, that the t distribution is quite similar to the standard normal distribution if the sample size n is large.

We can also cast the hypothesis test in terms of a **confidence interval**. The 95% confidence interval around the sample mean is the range of values within nearly two standard deviations ($s_\pi/\sqrt{n}$) of the mean. So the lower limit of the confidence interval is $\overline{\pi} - 1.96 s_\pi/\sqrt{n}$, and the upper limit is $\overline{\pi} + 1.96 s_\pi/\sqrt{n}$. We reject the null hypothesis that the population mean of economic profit is zero if the 95% confidence interval doesn't include zero.

Practice Question

In the (fictitious) franchise data, $\overline{\pi} = \$2{,}500$, $s_\pi = \$70{,}711$, and $n = 5{,}000$. What is the $t-$ratio is this sample? What is the 95% confidence interval? Does the data reject the hypothesis that economic profit is zero?

References

Wooldridge, Jeffrey M. *Introductory Econometrics: A Modern Approach*, 5/e. Cengage Learning, 2013.

Stock, James H., and Watson, Mark W. *Introduction to Econometrics*, 3/e. Pearson Education, 2015.

Acknowledgements

Shaoying Ma created this and other tutorials in the *Math for Economics* series. Kenneth McLaughlin supervised her work and edited the final product.