

Chapter 17. NORMAL DISTRIBUTIONS.

Random variables were introduced in Chapter 8 and normal distributions were defined. If the random variables X_1, X_2, \dots, X_n represent n successive and independent observations of a given random variable X , then the sample mean was defined to be the new random variable

$$\bar{x} = \frac{1}{n} \sum_i X_i .$$

We proved in Chapter 8 that

$$E_{\bar{x}} = E_X$$

and

$$V_{\bar{x}} = \frac{1}{n} V_X .$$

We also stated the Central Limit Theorem which asserts that for a random variable X possessing both mean and variance, the distribution of \bar{x} become more and more nearly normal as n increases.

Two further random variables can be associated with the above observation procedure. We define

$$S^2 = \sum_i (X_i - \bar{x})^2$$

and

$$s^2 = \frac{1}{n} S^2 .$$

S^2 is called the sum of squares and s^2 is called the sample variance. We also define

$$s = \sqrt{s^2}$$

which is called the sample standard deviation. In an appendix to this chapter, we show that

$$E_{s^2} = (n-1)V_X$$

and hence that

$$E_{s^2} = \frac{n-1}{n} V_X .$$

For the case where X is normal, we can also show that

$$V_{s^2} = \frac{2(n-1)}{n^2} (V_X)^2 ,$$

which suggests that as n increases, the distribution for the random variable s^2 becomes more and more closely concentrated about the expected value $\frac{n-1}{n} V_X$. Hence, in statistical applications, we can expect to use \bar{x} as an estimate for E_X and $\frac{n}{n-1} s^2$ as an estimate for V_X .

The random variables \bar{x} and s^2 are known as sampling variables because they are defined in terms of the sample X_1, \dots, X_n . Sampling variables such as \bar{x} and s^2 are important because they are convenient to use in defining various statistical metrics. The distributions of these sampling variables and metrics are called sampling distributions. Sampling distributions depend upon the distribution of the original random variable X . The results above, however, together with the Central Limit Theorem, show that when strong enough assumptions are made about the universe of models U , it is possible to reach conclusions about a sampling distribution even when the distribution of X is not fully known.

Much of the great early work in classical statistics can be described in the following way:

- (a) Certain useful sampling variables are defined.
- (b) A universe of models is assumed.
- (c) A metric is defined using the sampling variables.
- (d) The metric is shown to be well defined for certain appropriate composite models.
- (e) A sampling distribution for the metric is calculated.
- (f) This sampling distribution is used for DLS calculations and hence for hypothesis testing and confidence regions.

The special art of classical statisticians lay in the elegant ways in which they accomplished (c), (d), and (e).

Adding independent random variables. If we are given random variables X and Y , we can define a single experimental procedure yielding X and Y as independent random variables. (See Chapter 8.) We can then define $Z = X + Y$ as a new random variable associated with this procedure. Z is called the sum of the independent variables X and Y . The distribution for Z is then determined by (and can be found from) the distributions for X and Y . For example, if X and Y take values on the non-negative integers and p_i^X , p_j^Y , p_k^Z are the probability functions for X , Y , and Z , then for each n ,

$$p_n^Z = p_0^X p_n^Y + p_1^X p_{n-1}^Y + \dots + p_n^X p_0^Y.$$

Similarly, if X , Y , and Z are continuous random variables on the interval $(-\infty, \infty)$ and f^X , f^Y , f^Z are the density functions for X , Y , and Z , then it is easy to show that for each z ,

$$f^Z(z) = \int_{-\infty}^{\infty} f^X(t) f^Y(z-t) dt.$$

(This follows from the assumption of independence together with result (1) in the proofs at the end of Chapter 8, when we make the change of variables $z = z$ and $y = z-x$.) We shall look further at these operations in Chapter 19.

Using the integral formula above, we can show by direct integration that the sum of two independent normal variables must be a normal variable. Using the formula for discrete variables given, we can also show that the sum of two Poisson variables is a Poisson variable. In both cases, it is immediate from the facts given in Chapter 8, that the mean of the sum is the sum of the means and that the variance of the sum is the sum of the variances. A similar conclusion holds for the variable $Z = X - Y$, where X and Y are independent normal variables. In this case, Z must also be normal. Since $Z = X - Y = X + (-Y)$, we have $E_Z = E_X + E_{-Y} = E_X - E_Y$ and $V_Z = V_X + V_{-Y} = V_X + V_Y$. Thus, for example, if X is normal with mean 4 and standard deviation 3, and Y is normal with mean 5 and standard deviation 4, then $Z = X - Y$ must be normal with mean -1 and standard deviation 5. (Here $V_X = 3^2 = 9$, $V_Y = 4^2 = 16$, and hence $V_Z = 9 + 16 = 25$ giving $\sigma_Z = \sqrt{25} = 5$.)

In general, the sum of two independent binomial variables is not binomial, except in the special case where the individual success probability p is the same for the two given variables. Then (as is intuitively obvious) the sum is also a binomial variable with the same p and with a number of trials equal to the sum of the numbers of trials for the two given variables separately.

It is also possible to show, as we shall see in Chapter 19, that if X is normal, then, for every constant a , aX is normal. It follows that any linear combination of independent normal variables must itself be normal.

The Central Limit Theorem. The Central Limit Theorem was stated in Chapter 8 as a statement about the distribution of \bar{x} .

The Central Limit Theorem also holds in a more general form for sums of independent random variables $X_1 + X_2 + \dots + X_n$ where the variables X_1, X_2, \dots need not have the same distributions, but must have variances of about the same size. From an intuitive point of view, this general theorem says that the combined effect (as a random variable) of a large number of independent small effects must have a distribution that is approximately normal.

The proof of the Central Limit Theorem requires special analytic techniques to be indicated in Chapter 19.

Common occurrence of the normal distribution. The normal distribution appears frequently in applications of probability and statistics. Many random variables encountered in physical experiments prove to be approximately normal. (This is usually the case, for example, when the given random variable

is the value of a single direct observational measurement of some physical quantity such as length or temperature.) The Central Limit Theorem in its general form provides a conceptual explanation for the common occurrence of the normal distribution. If the observed value of a given random variable can be viewed as the combined effect of a large number of small independent factors, then the theorem tells us that the distribution should be normal. Thus certain random variables encountered in biology, such as height of a randomly chosen adult in some given human population, can be expected to be normal, because they can be viewed as the result of a large number of independent (and individually small) genetic and environmental factors. Similarly, the error in making a physical measurement can be expected (as a random variable) to be normal, if it can be viewed as the combined effect of a number of smaller, independent effects.

In statistical analysis, considerations of this kind often permit us to assume that our universe of models is a collection of normal distributions. When we make this assumption, certain stronger statistical methods become available to us, as we shall see below.

Standardized variables. Recall from Chapter 8 that if a random variable X has mean μ and variance σ^2 , then the new variable Y defined by

$$y = \frac{X - \mu}{\sigma}$$

is called the standardized form of X . Evidently, $E_y = 0$ and $V_y = 1$. If we know from assumption (or by the Central Limit

Theorem) that a variable X is normal, and if Y is the standardized form of X , then Y will have the standard normal curve as its density function, and tables of standard normal areas can be used to find probabilities for Y and hence for X .

An example of a standardized variable occurs in connection with normal approximation. Let X be the number of successes in a binomial experiment with n trials and success probability p . Then $E_X = np$, and $V_X = npq$ as we saw in Chapter 16. Hence the standardized form of X is

$$\frac{X-np}{\sqrt{npq}}$$

and this is the expression (with an added correction for bar width) that we used in normal approximation. Similarly, in normal approximation to the Poisson distribution, the standardized form $\frac{X-m}{\sqrt{m}}$ was used.

Statistical methods for normal distributions. We illustrate some of these in the examples that follow.

Let the random variable X be the height in inches of a student chosen at random from a given population of students. Assume that 100 independent observations of the random variable X are made, yielding $X_1 = 70$, $X_2 = 72.5$, ..., $X_{100} = 68.4$. Assume further that from these observations we calculate $\bar{x} = 70.5$ and $s^2 = 9.0$. What does this information tell us about the true distribution of X ? (We would know the true distribution exactly if we knew the heights of all students in the underlying population from which the random selection is made.)

There are several different approaches to this question, depending upon: (i) the information we assume to begin with about the true distribution (in other words, the universe of models that we assume); and (ii) whether we use a hypothesis test or a confidence region. We call the mean of the true distribution μ and the variance σ^2 .

Example 1. We assume that the true distribution is normal with known variance but unknown mean. In particular, let us assume that we know, from studies of other similar situations, that $\sigma^2 = 4$. Thus we are taking, as our universe of models, all normal distributions with $\sigma^2 = 4$. We can now use our observed data as follows.

Hypothesis test: Take the hypothesis $\mu = 71$. \bar{x} must be normal, since \bar{x} is a linear combination of X_1, \dots, X_n , and each of X_1, \dots, X_n is normal. \bar{x} has variance $= \frac{\sigma^2}{n}$ and hence standard deviation $= \sigma/\sqrt{n}$. Hence the quantity

$$\frac{\bar{x} - \mu}{\sigma/\sqrt{n}}$$

is normal with mean $= 0$ and standard deviation $= 1$. Thus if we take $\frac{|\bar{x} - \mu|}{\sigma/\sqrt{n}}$ as our metric, we can calculate a DLS by using the standard normal distribution. In particular, we get

$$\frac{|\bar{x}_0 - \mu|}{\sigma/\sqrt{n}} = \frac{|70.5 - 71|}{2/\sqrt{10}} = \frac{0.5}{0.2} = 2.5$$

Thus the DLS $= 1 - 2A(2.5) = 0.012$, and our observation leads us to reject the hypothesis $\mu = 71$ at critical level 0.05 .

Confidence region: Assume 95% confidence level. Then we want those values of μ for which the DLS ≥ 0.05 . This means (again

using the fact that for each μ , $\frac{\bar{x} - \mu}{\sigma/\sqrt{n}}$ is standard normal) that $\frac{|\bar{x} - \mu|}{\sigma/\sqrt{n}}$ must be ≤ 1.96 (as $1 - 2A(1.96) = 0.05$). Putting in the particular observed value of $\bar{x} = 70.5$ and the assumed value of $\sigma^2 = 4$, we get

$$\frac{|70.5 - \mu|}{2/\sqrt{10}} \leq 1.96$$

$$|70.5 - \mu| \leq 0.39 = 0.4.$$

Thus $\mu = 70.5 \pm 0.4$ are confidence limits and $70.1 \leq \mu \leq 70.9$ is our desired confidence region.

Example 2. We assume that the true distribution is not-necessarily-normal, with known variance but unknown mean. Here the method of Example 1 works exactly as before, provided that n is large enough so that, by the Central Limit Theorem, the distribution of \bar{x} must be approximately normal. It follows that for each μ , $\frac{\bar{x} - \mu}{\sigma/\sqrt{n}}$ must have a standard normal distribution. The calculations for hypothesis tests and for confidence regions are thus identical with the calculations in Example 1 and yield the same numerical results.

Example 3. We assume that the true distribution is not-necessarily-normal with unknown variance and unknown mean. Here we can proceed exactly as in Example 2, provided that n is also large enough to ensure that $\frac{n}{n-1} s^2$ is a good estimate of the unknown true variance σ^2 . (It is possible to obtain a formula for V_s^2 , This formula, which is rather complicated, shows that V_s^2 gets

small as n increases.)

We then simply replace σ by

$\frac{\sqrt{n}}{\sqrt{n-1}} s$, or, equivalently, $\frac{\sigma}{\sqrt{n}}$ by $\frac{s}{\sqrt{n-1}}$. The calculations in the particular example above go as follows.

Hypothesis test: Take the hypothesis $\mu = 71$. Then $\frac{|\bar{x}-\mu|}{s/\sqrt{n-1}} = \frac{|70.5 - 71|}{3/\sqrt{99}} = 1.66$. Thus the DLS = $1 - 2A(1.66) = 0.10$, and the observation leads us to accept the hypothesis $\mu = 71$ at critical level 0.05.

Confidence region:

$$\frac{|\bar{x}-\mu|}{s/\sqrt{n-1}} = \frac{|70.5 - \mu|}{3/\sqrt{99}} \leq 1.96$$

$$|70.5 - \mu| \leq 0.59$$

Thus $\mu = 70.5 \pm 0.6$ are confidence limits, and

$$69.9 \leq \mu \leq 71.1$$

is the desired confidence region.

The basic formula from which we calculate all three of the above examples, and for both hypothesis tests and confidence regions, can be written (for critical level 0.05) as follows:

$$\frac{|\bar{x}-\mu|}{\sigma/\sqrt{n}} = 1.96 \quad \text{when } \sigma \text{ is known;}$$

$$\text{and} \quad \frac{|\bar{x}-\mu|}{s/\sqrt{n-1}} = 1.96 \quad \text{when } \sigma \text{ is unknown.}$$

(Here, as usual, we define $\sigma = \sqrt{\sigma^2}$ and $s = \sqrt{s^2}$.)


For Examples 2 and 3, n must be large enough to ensure that the Central Limit Theorem, and the desired approximation of σ^2 by $\frac{n}{n-1} s^2$, both hold.

The following case remains to be considered.

Example 4. We assume that the true distribution is normal with unknown variance and unknown mean. When n is large enough to ensure the desired approximations, this is simply a special case of Example 3, and the same formulas can be used. What if n is small? For example, what if we measure only five students and get $X_1 = 72$, $X_2 = 71.5$, $X_3 = 68$, $X_4 = 71$, $X_5 = 72.2$? Then $\bar{x} = 70.9$ and $s^2 = 2.3$. Can we carry out a hypothesis test or find a confidence region? We can do so as follows. We take our universe of models to be the set of all normal distributions. Within this universe, we define, for each value of μ , the composite model M_μ as follows: M_μ is the set of all normal distributions with mean μ . It can now be proved that for any given μ , the metric $\frac{\bar{x} - \mu}{s/\sqrt{n-1}}$ is well-defined on M_μ . We do not give the proof here. (It uses the methods of Chapters 12 and 19.) We call this metric Student's metric. The DLS of an observation (using Student's metric) can be obtained by using a certain standard density function known as the t-distribution with $n-1$ degrees of freedom. Like the chi-square curves, the t-distribution curves form a family of density functions, with a different curve $f_d(t)$ for each integer value $d > 0$, where d is called the number of degrees of freedom. The t-distribution curves are symmetrical, have mean = 0, and are rather similar in general shape to the standard

normal curve. As d increases, the curve $f_d(t)$ approaches more and more closely to the standard normal curve. Values for area under the t -curves are given in tables similar to the tables previously given for the chi-square curves, except that in the table given below, the area associated with a value t is the area lying under the curve between $-t$ and t . Areas are given in the top line, and t -values corresponding to those areas are given in the body of the table.

t - TABLES

Degrees of freedom	One-sample size†	Two-sided probability level 				
		.50	.80	.90	.95	.99
1	2	1.00	3.08	6.31	12.71	31.82
2	3	.82	1.89	2.92	4.30	6.96
3	4	.76	1.64	2.35	3.18	4.54
4	5	.74	1.53	2.13	2.78	3.75
5	6	.73	1.48	2.02	2.57	3.36
6	7	.72	1.44	1.94	2.45	3.14
7	8	.71	1.41	1.89	2.36	3.00
8	9	.71	1.40	1.86	2.31	2.90
9	10	.70	1.38	1.83	2.26	2.82
10	11	.70	1.37	1.81	2.23	2.76
15	16	.69	1.34	1.75	2.13	2.60
30	31	.68	1.31	1.70	2.04	2.46
50	51	.68	1.30	1.68	2.01	2.40
100	101	.68	1.29	1.66	1.98	2.37
1000	1001	.67	1.28	1.65	1.96	2.33
∞ ‡	∞ ‡	.67	1.28	1.64	1.96	2.33

† For setting confidence limits on the mean of a single sample.

‡ Standard normal distribution.

Thus for 4 degrees of freedom, area 0.95 is given by the t -value 2.78.

We can now perform hypothesis tests and find confidence regions as follows.

Hypothesis test: Take the hypothesis $\mu = 72.5$. Using $\bar{x} = 70.9$ and $s^2 = 2.3$, we get

$$\frac{|\bar{x}-\mu|}{s/\sqrt{n-1}} = \frac{|70.9-72.5|}{\sqrt{2.3}/\sqrt{4}} = \frac{1.6}{0.76} = 2.11$$

From the table, using $n-1 = 4$ degrees of freedom, we see that the area for the t -value 2.1 must fall between 0.80 and 0.90. Hence the DLS of the observation (given by the remaining area) must lie between 0.1 and 0.2. At the critical level 0.05, we would continue to accept the hypothesis.

Confidence region: For a 95% confidence region, we want those values of μ which give a DLS > 0.05 ; in other words, we want the values of μ whose t -values give an area < 0.95 . For $d = 4$, area 0.95 occurs at the t -value 2.78. Hence we have

$$\frac{|\bar{x}-\mu|}{s/\sqrt{n-1}} = \frac{|70.9-\mu|}{\sqrt{2.3}/\sqrt{4}} = 2.78$$

$$\text{Thus, } |70.9-\mu| = 2.11 = 2.1$$

So the confidence region, in this case, is

$$68.8 \leq \mu \leq 73.0.$$

The difference of two means. X and Y are two given random variables. We make n independent observations of X (represented by the random variables X_1, \dots, X_n) and m independent observations of Y (represented by the random variables Y_1, \dots, Y_m). A common form of statistical problem is

the following: to decide, on the basis of such observations, whether it is reasonable to conclude that X and Y have the same distribution. (This problem arises frequently in experimental studies, where we seek to determine if a treated group of subjects is different from a control group.) In Chapter 16, the WMW-metric gave us a non-parametric approach to problems of this kind. We now return to this problem, making the additional assumption that the distributions for X and Y are normal. We shall see that the assumption of this additional information leads us (as we might expect) to more powerful statistical methods of a parametric nature. We give three examples.

Example 5. We assume as hypothesis that the distributions for X and Y are the same distribution, that their common variance σ^2 is already known, but that their common mean is unknown. We take, as a composite model, all pairs of normal distributions where both members of each pair have the same specified variance σ^2 , and equal means. Then for any pair of mean μ , \bar{x} must be normal with mean μ and variance σ^2/n , and \bar{y} must be normal with mean μ and variance σ^2/m . Hence the random variable $\bar{x} - \bar{y}$ must be normal with mean 0 and variance $\frac{\sigma^2}{n} + \frac{\sigma^2}{m}$. Hence the random variable

$$\frac{\bar{x} - \bar{y}}{\sqrt{\sigma^2 \left(\frac{1}{n} + \frac{1}{m} \right)}}$$
 must be standard normal. This random variable can

then be used as a metric to measure how far an observation is from giving strongest confirmation that X and Y have the same distribution.

Example 6. We assume as hypothesis that the distributions for X and Y are the same normal distribution, but that their common variance and common mean are both unknown. Then, if n and m are sufficiently large, the observation can be used to estimate the unknown variance σ^2 as follows. Let S_1^2 be the sum of squares for the observations of X and let S_2^2 be the sum of squares for the observations of Y . It follows from the results of Chapter 8 that

$$E(S_1^2 + S_2^2) = E(S_1^2) + E(S_2^2) = (n-1)\sigma^2 + (m-1)\sigma^2 = (n+m-2)\sigma^2,$$

where σ^2 is the unknown common variance. Then

$$\hat{\sigma}^2 = \frac{S_1^2 + S_2^2}{n+m-2}$$

can be used as an estimate of σ^2 . Substituting this estimate for σ^2 in Example 5, we can complete our analysis, using the standard normal curve, exactly as in that example.

Example 7. For a final example, we consider the case where the distributions for X and Y are assumed (as hypothesis) to have the same normal distribution with common mean and common variance both unknown, and where n and m are not large. In this case, it can be shown that the random variable

$$\frac{\bar{x} - \bar{y}}{\sqrt{\hat{\sigma}^2 \left(\frac{1}{n} + \frac{1}{m} \right)}}$$

follows the t -distribution with $n + m - 2$ degrees of freedom. (See Chapter 19.) Using this random variable as a metric (to measure

confirmation of the composite null hypothesis that X and Y have the same mean), we have the t-test for the difference of two means, one of the most commonly used techniques in classical statistical parametric analysis.