

Math107: The Normal
Distribution and Hypothesis
Testing Using the Central Limit
Theorem

Dr. Richard Mikula

Fall 2009

The Standard Normal Distribution:

The standard normal or Gaussian distribution is the probability distribution for a particular continuous random variable Z , whose expected value is

$$E(Z) = 0$$

and standard deviation is

$$sd(Z) = 1.$$

The probability

$$P(a \leq Z \leq b)$$

is given by the area under the curve given by the graph of the function

$$y = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2},$$

the so-called bell curve, between the two vertical lines $x = a$ and $x = b$, and above the x axis (the horizontal axis).

Note:

$$e = 2.71828182846 \dots$$

$$= \frac{1}{0!} + \frac{1}{1!} + \frac{1}{2!} + \frac{1}{3!} + \frac{1}{4!} + \frac{1}{5!} + \frac{1}{6!} + \frac{1}{7!} + \dots$$

is an irrational number. Also

$$e^k$$

represents e raised to the k th power, for any number k .

The so-called **bell curve** is symmetric about the vertical axis (the y axis).* We need to keep this in mind in calculations we will perform using a table that is to follow.

*We will see a graph of this curve in lecture.

To find the probabilities

$$P(a \leq Z \leq b)$$

for a random variable that has the standard normal distribution, we need to use a table*, which is in the text. Below is a version of this table.

*Note that the probabilities $P(a \leq Z \leq b)$, $P(a \leq Z < b)$, $P(a < Z \leq b)$, $P(a < Z < b)$ are all the same since Z is a continuous random variable, where here a, b are any fixed numbers.

Standard Normal Distribution Table:

The below table gives the probabilities

$$P(0 \leq Z \leq z)$$

where the z value here is the independent variable.

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.000	.004	.008	.012	.016	.029	.024	.028	.032	.036
0.1	.040	.044	.048	.052	.056	.060	.064	.068	.071	.075
0.2	.079	.083	.087	.091	.095	.099	.103	.106	.110	.114
0.3	.118	.122	.126	.129	.133	.137	.141	.144	.148	.152
0.4	.155	.159	.163	.166	.170	.174	.177	.181	.184	.190
0.5	.192	.195	.199	.202	.205	.209	.212	.216	.219	.222
0.6	.226	.229	.232	.236	.239	.242	.245	.249	.252	.255
0.7	.258	.261	.264	.267	.270	.273	.276	.279	.282	.285
0.8	.288	.291	.294	.297	.300	.302	.305	.308	.311	.313
0.9	.316	.319	.321	.324	.326	.329	.332	.334	.337	.339

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
1.0	.341	.345	.346	.349	.351	.353	.355	.358	.360	.362
1.1	.364	.367	.369	.371	.373	.375	.377	.379	.381	.383
1.2	.385	.387	.389	.391	.393	.394	.396	.398	.400	.402
1.3	.403	.405	.407	.408	.410	.412	.413	.415	.416	.418
1.4	.419	.421	.422	.424	.425	.427	.428	.429	.431	.432
1.5	.433	.435	.436	.437	.438	.439	.441	.442	.443	.444
1.6	.445	.446	.447	.448	.450	.451	.452	.453	.454	.455
1.7	.455	.456	.457	.458	.459	.460	.461	.462	.463	.463
1.8	.464	.465	.466	.466	.467	.468	.469	.469	.470	.471
1.9	.471	.472	.473	.473	.474	.474	.475	.476	.476	.477

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
2.0	.477	.478	.478	.479	.479	.480	.480	.481	.481	.482
2.1	.482	.483	.483	.483	.484	.484	.485	.485	.485	.486
2.2	.486	.486	.487	.487	.488	.488	.488	.488	.489	.489
2.3	.489	.490	.490	.490	.490	.491	.491	.491	.491	.492
2.4	.492	.492	.492	.493	.493	.493	.493	.493	.493	.494
2.5	.494	.494	.494	.494	.495	.495	.495	.495	.495	.495
2.6	.495	.496	.496	.496	.496	.496	.496	.496	.496	.496
2.7	.497	.497	.497	.497	.497	.497	.497	.497	.497	.497
2.8	.497	.498	.498	.498	.498	.498	.498	.498	.498	.498
2.9	.498	.498	.498	.498	.498	.498	.499	.499	.499	.499

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
3.0	.499	.499	.499	.499	.499	.499	.499	.499	.499	.499
3.1	.499	.499	.499	.499	.499	.499	.499	.499	.499	.499
3.2	.499	.499	.499	.499	.499	.499	.499	.499	.499	.499
3.3	.499	.499	.499	.499	.499	.499	.499	.500	.500	.500

Moreover, for $z \geq 3.4$ we have

$$P(0 \leq Z \leq z) = 0.5$$

as well. Also, we will not distinguish between $<$ and \leq , or $>$ and \geq .

How to use the above table:

Let $a, b \geq 0$. Then we have the following:

$$1. P(Z \leq -a) = 0.5 - P(0 \leq Z \leq a)$$

$$2. P(-a \leq Z \leq -b) = P(0 \leq Z \leq a) - P(0 \leq Z \leq b)$$

$$3. P(-a \leq Z \leq 0) = P(0 \leq Z \leq a)$$

$$4. P(-a \leq Z \leq b) = P(0 \leq Z \leq a) + P(0 \leq Z \leq b)$$

5. $P(0 \leq Z \leq b)$ is in the above table.

$$6. P(a \leq Z \leq b) = P(0 \leq Z \leq b) - P(0 \leq Z \leq a)$$

$$7. P(Z \geq b) = 0.5 - P(0 \leq Z \leq b)$$

$$8. P(-a \leq Z) = 0.5 + P(0 \leq Z \leq a)$$

$$9. P(Z \leq b) = 0.5 + P(0 \leq Z \leq b)$$

For example, if we are told that a random variable Z for some experiment had the standard normal distribution, we could calculate the following probabilities

1. $P(-\infty < Z < -0.56)$

2. $P(Z < -0.572)$

3. $P(-3.5 \leq Z < -0.5)$

4. $P(0.5 < Z < 3.5)$

5. $P(Z > 2.5)$

6. $P(-1.5 < Z < 2.5)$

7. $P(-1 < Z < 1)$

8. $P(-2 < Z < 2)$

9. $P(-3 < Z < 3)$

10. $P(-4 < Z < 4)$

Answers:

1.

$$\begin{aligned} P(Z < -0.56) &= 0.5 - P(0 \leq Z \leq 0.56) \\ &= 0.5 - 0.212 = 0.288. \end{aligned}$$

2.

$$\begin{aligned} P(Z < -0.572) &= 0.5 - P(0 \leq Z \leq 0.572) \\ &\approx 0.5 - P(0 \leq Z \leq 0.57) = 0.5 - 0.215 = 0.284. \end{aligned}$$

3.

$$\begin{aligned} &P(-3.5 \leq Z < -0.5) \\ &= P(0 \leq Z \leq 3.5) - P(0 \leq Z \leq 0.5) \\ &= 0.500 - 0.192 = 0.308. \end{aligned}$$

4.

$$\begin{aligned} & P(0.5 < Z < 3.5) \\ &= P(0 \leq Z \leq 3.5) - P(0 \leq Z \leq 0.5) \\ &= 0.500 - 0.192 = 0.308. \end{aligned}$$

5.

$$P(Z > 2.5) = 0.5 - P(0 \leq Z \leq 2.5) = 0.006$$

6.

$$\begin{aligned} & P(-1.5 < Z < 2.5) \\ &= P(0 \leq Z \leq 2.5) + P(0 \leq Z \leq 1.5) \\ &= 0.494 + 0.433 = 0.927 \end{aligned}$$

7.

$$P(-1 < Z < 1) = P(0 \leq Z \leq 1) + P(0 \leq Z \leq 1)$$

$$= 2 \cdot 0.341 = 0.682$$

8.

$$\begin{aligned} P(-2 < Z < 2) &= P(0 \leq Z \leq 2) + P(0 \leq Z \leq 2) \\ &= 2 \cdot 0.477 = 0.954 \end{aligned}$$

9.

$$\begin{aligned} P(-3 < Z < 3) &= P(0 \leq Z \leq 3) + P(0 \leq Z \leq 3) \\ &= 2 \cdot 0.499 = 0.988 \end{aligned}$$

10.

$$\begin{aligned} P(-4 < Z < 4) &= P(0 \leq Z \leq 4) + P(0 \leq Z \leq 4) \\ &= 2 \cdot 0.5 = 1 \end{aligned}$$

Let us now look at an example:

The **Wechsler IQ test** is an intelligence test whose scores have a normal distribution with mean 100 and standard deviation 15. If X represents the IQ score of a randomly chosen person, then the quantity

$$Z = \frac{X - 100}{15}$$

has a standard normal distribution. That is

$$E(Z) = 0, \quad sd(Z) = 1$$

and Z has a normal distribution.

Suppose a person is selected at random. Let X represent the IQ score of such a person, find the following probabilities:

1. $P(X \geq 135)$

2. $P(80 \leq X \leq 120)$

3. $P(X < 70)$

4. $P(X > 150)$

Solutions:

We compute

$$Z = \frac{X - 100}{15},$$

and observe that Z has the standard normal distribution with mean and standard deviation

$$E(Z) = 0, \quad sd(Z) = 1.$$

Thus:

1.

$$\begin{aligned} P(X \geq 135) &= P\left(\frac{X - 100}{15} \geq \frac{135 - 100}{15}\right) \\ &\approx P(Z \geq 2.33) \end{aligned}$$

$$= 0.5 - P(0 \leq Z \leq 2.33) = 0.5 - 0.490 = 0.01$$

Thus, people with IQ's over 135 are in the top 1 percentile of the population with respect to IQ scores.

2.

$$\begin{aligned} & P(80 \leq X \leq 120) \\ &= P\left(\frac{80 - 100}{15} \leq \frac{X - 100}{15} \leq \frac{120 - 100}{15}\right) \\ &\approx P(-1.33 \leq Z \leq 1.33) \\ &P(0 \leq Z \leq 1.33) + P(0 \leq Z \leq 1.33) \\ &= 2 \cdot 0.408 = 0.816 \end{aligned}$$

Thus, we see that roughly 82 percent of the population have their IQ scores lie in the range of 80 to 120.

3.

$$\begin{aligned} P(X < 70) &= P\left(\frac{X - 100}{15} < \frac{70 - 100}{15}\right) \\ &= P(Z < -2) = 0.5 - P(0 \leq Z \leq 2) \\ &= 0.5 - 0.477 = 0.023 \end{aligned}$$

4.

$$\begin{aligned} P(X > 150) &= P\left(\frac{X - 100}{15} > \frac{150 - 100}{15}\right) \\ &\approx P(Z > 3.33) = 0.5 - P(0 \leq Z \leq 3.33) = 0.001 \end{aligned}$$

Understanding Data from Samples:

Interpretation of data is the essential ingredient in making predictions about the population in question and in the quantification of the amount of uncertainty in these predictions.

Suppose X is a Random Variable with mean $E(X) = \mu$ and standard deviation $sd(X) = \sigma$, defined on some population. A Random Sample of size n , (drawn with replacement,) yields n values,

$$X_1, X_2, \dots, X_n$$

for X . These values are independent of one another. Thus they may be considered as n independent random variables, identically distributed with the same mean μ and standard deviation σ .

We define the so-called *sample mean* to be

$$\bar{X} = \frac{X_1 + X_2 + \cdots + X_n}{n}.$$

We may think of \bar{X} as a random variable, then from properties of the mean (expectation) and standard deviation of linear combinations of random variables, we may see that

$$E(\bar{X}) = \mu, \quad \text{and} \quad sd(\bar{X}) = \frac{\sigma}{\sqrt{n}}.$$

One very important result about the distribution of the sample mean is the so-called **Central Limit Theorem**:

Suppose X is a random variable with mean $E(X) = \mu$ and standard deviation $sd(X) = \sigma > 0$ defined on some population. If n is large (which in practice $n \geq 30$ is sufficient), then the sample mean is approximately normally distributed with mean $E(\bar{X}) = \mu$ and standard deviation $sd(\bar{X}) = \frac{\sigma}{\sqrt{n}}$.

Note: If $n < 30$, the normal approximation of \bar{X} is only good if X is (approximately) normally distributed itself.

More on Normally Distributed Populations and the Central Limit Theorem.

We have that the Wechsler I.Q. scores of people in the U.S. have a mean μ of 100 and a standard deviation σ of 15.

Suppose we were to randomly sample 30 people, and administer the I.Q. test; let

$$X_1, X_2, X_3, \dots, X_{29}, X_{30}$$

be their scores. We define

$$\bar{X} = \frac{X_1 + X_2 + \dots + X_{30}}{30}.$$

Since the X_i 's can be anything, it is reasonable to expect that the \bar{X} can be anything.

We consider the following questions:

- How reasonable is it to expect \bar{X} to be close to 100?
- How close should it be to 100?

By the central limit theorem, we know that \bar{X} * is approximately normally distributed, with mean $E(\bar{X}) = 100$ and standard deviation $sd(\bar{X}) = \frac{15}{\sqrt{30}} \approx 2.7386$.

*That is, we think of \bar{X} as a random variable via taking different samples of 30 people and calculating their average I.Q. score $\bar{X} = \frac{X_1 + \dots + X_{30}}{30}$.

Hypothesis Testing:

When studying a population, we would like to make claims about population parameters. It is impractical in many instances to study the entire population, so we decide to take random samples from the population, and study these instead. From the parameters we observe from our samples, we wish to make statements about the entire population.

Recall our previous example. Suppose we took a poll of 30 people and let their I.Q. scores be given by

65, 70, 78, 80, 85, 90, 95, 97, 98, 98

98, 98, 99, 99, 99, 100, 100, 101, 102, 103

103, 104, 104, 105, 106, 108, 115, 125, 140, 145.

Computing the sample mean we get

$$\bar{X} = \frac{3010}{30} = 100.\bar{3}.$$

If we wish to test whether or not the population mean is actually 100, we should see what the probability of observing a sample mean of $100.\bar{3}$ is. We have either H_0 : that the population mean is 100, or H_1 : that the population mean is not 100. We call H_0 our **null hypothesis** and H_1 our **alternate hypothesis**. Clearly one of these is true.

The Central Limit Theorem tells us that \bar{X} is approximately normally distributed with a mean of 100 and a standard deviation of $\frac{15}{\sqrt{30}} \approx 2.7386$. We will use this as a gauge to see the likelihood of observing a sample mean of $100.\bar{3}$.

When we test to see if we reject the null hypothesis, or do not reject the null hypothesis, we set a **level of significance** α . We will use in this case $\alpha = 0.05$. If we observe a result that has a probability of no more than α of occurring, assuming the null hypothesis holds, we will reject the null hypothesis. Otherwise, we choose not to reject the null hypothesis.

By the Central Limit Theorem, we know that \bar{X} is approximately normally distributed with $E(\bar{X}) = 100$ and $sd(\bar{X}) = \frac{15}{\sqrt{30}} \approx 2.7386$. We examine the region symmetric about the mean which has a probability of 0.95 of happening. By our Z -table we see that

$$P(-1.96 \leq Z \leq 1.96) = 0.95.$$

We then let

$$Z = \frac{\bar{X} - 100}{\frac{15}{\sqrt{30}}}$$

and see that this is equivalent to

$$P(94.6323 \leq \bar{X} \leq 105.3677) = 0.95.$$

Since our observed value of \bar{X} lies in this region that has a probability of $0.95 = 1 - \alpha$ of occurring, we do not reject the null hypothesis.

Example: Previous research indicates that a population mean is $\mu = 50$ with a standard deviation of $\sigma = 3$.

A random sample of 36 measurements yields $\bar{X} = 48$. Does this sample provide sufficient evidence at the $\alpha = 0.05$ level of significance that our population mean is not 50?

Solution: Our null hypothesis is $H_0: \mu = 50$ and our alternate hypothesis is $H_1: \mu \neq 50$.

As above, we observe that

$$P(-1.96 \leq Z \leq 1.96) = 0.95.$$

Letting $Z = \frac{\bar{X} - 50}{\frac{3}{\sqrt{36}}}$, we see that when $\bar{X} = 48$, $Z = -4$. Since $-4 \notin [-1.96, 1.96]$, we reject the null hypothesis in favor of the alternate hypothesis.

In the previous two examples, The region which corresponds with $-1.96 \leq Z \leq 1.96$ is called the **95 percent confidence interval** for μ . If our observed value of \bar{X} had $\frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}}$ lie in this interval, we chose not to reject the null hypothesis. Otherwise, we rejected the null hypothesis in favor of the alternate hypothesis.

Home Work Exercises:

- Suppose that a random sample of 100 students at a particular university had a sample mean I.Q. score of 110. Does this evidence support the claim that the universities students have an average I.Q. score of 105 with a standard deviation of 12?
- Suppose old weather records for a particular date suggest that the average high temperature for that date is 22 degrees Celsius, with a standard deviation of 2 degrees Celsius. However, suppose you have checked the statistics for the past 49 years, and you discovered an average of 24 degrees for that date from these 49 recorded temperatures. Does this evidence seem to support the claim of the old weather records?

Clearly the normal distribution is an important probability distribution. We see that sample means are approximately normally distributed, if the sample size is sufficiently large (at least 30).

Moreover, the binomial distribution may be approximated by a normal distribution if $np, n(1 - p) \geq 5$.

Moreover, if X is the number of successes for a sequence of Bernoulli trials with $np, nq \geq 5$ we approximate $P(X = k)$ by

$$P(X = k) \approx P(k - 0.5 \leq Y \leq k + 0.5)$$

where the random variable Y has a normal distribution with

$$E(Y) = np, \quad sd(Y) = \sqrt{npq}.$$

Hence, we may wish to see if a population appears to be normally distributed. To do this suppose we have a random sample from the population

$$X_1, X_2, \dots, X_{n-1}, X_n.$$

Suppose the the X_i 's are ordered in increasing order. Then give the data value a percentile ranking. For the value X_i we use $\frac{i-\frac{1}{2}}{n}$ as the ranking. We then find the Z score Z_i that has the property that

$$P(Z \leq Z_i) \approx \frac{i - \frac{1}{2}}{n}.$$

We then test the strength of the linear correlation between the original data $\{X_1, X_2, \dots, X_n\}$ and the corresponding $\{Z_1, Z_2, \dots, Z_n\}$ scores.

Example: Does data set

95.5, 80.8, 103.7, 119.1, 118,

126, 67.2, 96.5, 116.4, 83.7

appear to be normally distributed?

Solution:

Data	Rank	percentile	<i>Z</i> -score
67.2	1	0.05	-1.64
80.8	2	0.15	-1.04
83.7	3	0.25	-0.67
95.5	4	0.35	-0.39
96.5	5	0.45	-0.13
103.7	6	0.55	0.13
116.4	7	0.65	0.39
118	8	0.75	0.67
119.1	9	0.85	1.04
126	10	0.95	1.64

In this example, the Pearson Correlation Coefficient is 0.976. Thus the data set is probably normally distributed.

Homework: Suppose that the following are I.Q. scores for 10 random people

85, 95, 98, 99, 101, 101, 102, 102, 104, 120.

Do they appear to be normally distributed? Please justify your answer.

Confidence Intervals, revisited

In the earlier stuff we discussed on hypothesis testing, we made the assumption that the standard deviation was known, but we tested whether or not the mean is a certain value.

The basis of our **confidence interval** procedure is: If X has mean μ and standard deviation σ , then for $n \geq 30$ we know via the central limit theorem that \bar{X} is approximately normally distributed with mean μ and standard deviation $\frac{\sigma}{\sqrt{n}}$.

Thus, we can say that $100(1 - \alpha)$ percent of all samples of size n have means within the interval

$$\left[\mu - Z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}, \mu + Z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \right]$$

where here $Z_{\frac{\alpha}{2}}$ is the Z -score that has the area under the bell curve to the right of Z being $\frac{\alpha}{2}$.

Equivalently, we can say that $100(1 - \alpha)$ percent of all samples of size n have the property that the interval

$$\left[\bar{X} - Z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}, \bar{X} + Z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \right]$$

contains μ .

Margin of Error:

In the above discussion, we see that the **margin of error** for our estimate of μ is

$$e = Z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}.$$

Thus, we see that the sample size required for a $(1 - \alpha)$ -level confidence interval for μ with a specified margin of error e , is given by the formula

$$n = \left(Z_{\frac{\alpha}{2}} \frac{\sigma}{e} \right)^2$$

rounded up to the nearest integer.

Example: Suppose we wish to determine the mean age μ of the civilian labor force. Suppose that the standard deviation $\sigma = 12.1$ years is known. **a:** Determine the sample size n needed in order to be 95 percent confident that μ is within 0.5 year of the estimate \bar{X} . **b:** Find a 95 percent confidence interval for μ if a sample of the size found in part **a** has a mean age of 38.8 years.

Solutions: a: Here $\sigma = 12.1$ and the margin of error $e = 0.5$, thus

$$n = \left(1.96 \frac{12.1}{0.5}\right)^2 = 2249.79.$$

Thus, if 2250 people in the civilian labor force are randomly selected, we can be 95 percent confident that the mean age of all people in the civilian labor force is within 0.5 years of the mean sample age.

b: Here we use $\alpha = 0.05$, $\sigma = 12.1$, $\bar{X} = 38.8$ and $n = 2250$ to get the confidence interval

$$\left[38.8 - 1.96 \frac{12.1}{\sqrt{2250}}, 38.8 + 1.96 \cdot \frac{12.1}{\sqrt{2250}}\right]$$

or simply

$$[38.3, 39.3].$$

Homework:

In estimating the mean monthly fuel expenditure, μ , per household vehicle, the U.S. Energy Information Administration takes a sample of size 6841. Assuming $\sigma = 20.65$ dollars, determine the margin or error in estimating μ at the 95 percent level of confidence. **Answer:** 0.49 dollars.

In all the examples we have thus considered we have studied a population which we wanted to determine a confidence interval for μ , the mean of the population, assuming that the standard deviation σ is known. For a sample of size n , the random variable \bar{X} is approximately normally distributed. Hence

$$Z = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

has the standard normal distribution.

Suppose that the population standard deviation σ is unknown, which is probably the case in practice. The best we can do is also estimate the standard deviation σ by the sample standard deviation s . We then base our confidence interval procedure on the resulting variable

$$T = \frac{\bar{X} - \mu}{\frac{s}{\sqrt{n}}}.$$

Unfortunately, the random variable T does not have a normal distribution. The distribution of T is the so-called **Student's t distribution**.