SAMPLING DISTRIBUTIONS & INFERENCES استنتاج CONCERNING A MEAN

Populations and Samples

If a population is infinite, it is impossible to observe all its values, and even if it is finite it may be impractical or uneconomical to observe it in its entirety. Thus, it is usually necessary to use a **sample**, a part of a population, and deduce from it results refer to the entire population. Clearly, such results can be useful only if the sample is in some way "representative" of the population.

The Sampling Distribution of the Mean (σ known)

Theorem 1 If a random sample of size n is taken from a population having the mean μ and the variance σ^2, then $\overline{X}$ is a random variable whose distribution has the mean μ.

For samples from infinite populations the variance of this distribution is $\frac{\sigma^2}{n}$

For samples from a finite population of size N the variance is $\frac{\sigma^2}{N}\ \frac{N-n}{N-1}$ where $\frac{N-n}{N-1}$ is often called the **finite population correction factor.**

Theorem 2 If $\overline{X}$ is the mean of a random sample of size n taken from a population having the mean μ and the finite variance σ^2, then

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

is a random variable whose distribution function approaches that of the standard normal distributions as $n \to \infty$.

Example 1: Car mufflers are constructed by nearly automatic machines. One manufacturer finds that, for any type of car muffler, the time for a person to set up and complete a production run has **a normal distribution with mean 1.82** hours and **standard deviation 1.20**. What is the probability that the sample mean of the next **40** runs will be from **1.65 to 2.04** hours.

Solution:

According to **Theorem 2**

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} = \frac{1.65 - 1.82}{1.2/\sqrt{40}} = -0.896$$

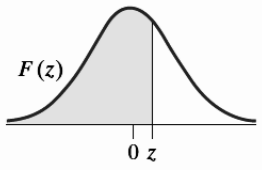
$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} = \frac{2.04 - 1.82}{1.2/\sqrt{40}} = 1.16$$

From Table 3

$$P(z = 1.16) - P(z = -.896) = 0.877 - \left(\frac{0.1867 + 0.1841}{2}\right) = 0.6916$$

Table 3

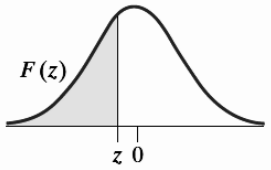
$$F(z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} e^{-t^2/2}\, dt$$



z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177

Table 3 Standard Normal Distribution Function

$$F(z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} e^{-t^2/2}\, dt$$



z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
−0.9	0.1841	0.1814	0.1788	0.1762	0.1736	0.1711	0.1685	0.1660	0.1635	0.1611
−0.8	0.2119	0.2090	0.2061	0.2033	0.2005	0.1977	0.1949	0.1922	0.1894	0.1867
−0.7	0.2420	0.2389	0.2358	0.2327	0.2296	0.2266	0.2236	0.2206	0.2177	0.2148

The Sampling Distribution of the Mean (σ unknown)

Theorem 3 If $\bar{X}$ is the mean of a random sample of size n taken from a normal population having the mean μ and the variance σ^2, and

$$S^2 = \sum_{i=1}^{n} \frac{(X_i - \bar{X})^2}{n-1}$$

Then

$$t = \frac{\bar{X} - \mu}{S/\sqrt{n}}$$

is a random variable having the t distribution with the parameter $\nu = n - 1$ given in Table 4.

INFERENCES CONCERNING A MEAN
استنتاجات المتوسط

Statistical Approaches to Making Generalizations

Suppose that a random sample of n observations, from some population, leads. As matter of fact, to obtain new knowledge about a process or phenomena, appropriate data must be collected. Usually, it is not possible to obtain a complete set of data but only a sample. Statistical inference الاستدلال الاحصائي developed whenever it is needed to make **generalizations** تعميم about a population on the basis of a sample. It has to be mentioned that the generalization is usually called a **statistical inference** or just an **inference**.

The first step in making a good statistical inference is to model the population. Next, any statistic parameter, such as $\overline{X}$ or S^2, are calculated as a function of the sample. There are two main steps of statistical inference to **estimation of population parameters** معاملات المجتمع and **testing hypotheses**. First **estimation** can be either ***a point estimator*** that gives a single number estimate of the value of the parameter or ***an interval estimate*** that specifies an interval of reasonable values for the parameter. **A test of hypotheses** provides the answer to whether the data support or deny an investigator's claim about the value of the parameter.

Point Estimation

Basically, **point estimation** concerns the choosing of a statistic, that is, a single number calculated from sample data. This property suggests considering the sample mean $\bar{X}$ as a point estimator of the population mean μ. In the context of point estimation, the quantity of **standard error** can be calculated as $\frac{S}{\sqrt{n}}$

Example 2: Of all the waste materials entering landfills النفايات, a substantial proportion consists of construction and demolition materials. From the standpoint of green engineering, before incorporating these materials into the base for new or rehabilitated roadways, engineers must assess their strength. Generally, higher values imply a stiffer base which increases pavement life. Measurements of the elasticity modulus (MPa) on $n = 18$ specimens of recycled concrete aggregate produce the ordered values

136 143 147 151 158 160
161 163 165 167 173 174
181 181 185 188 190 205

The descriptive summary for the sample are **sample mean** $\bar{x} = 168.2$ and **sample standard deviation** $S = 18.10$

It is required to estimated standard error

Solution:

Our **point estimator** is sample mean is $\bar{x} = 168.2$ and S **= 18.10 MPa** is the calculated sample standard deviation. The estimated standard error is $\frac{S}{\sqrt{n}} = \frac{18.1}{\sqrt{18}} = 4.27$

Maximum Error of Estimate with High Probability

When a sample mean is used to estimate the mean of a population, it is no doubt that the chances are almost nonexistent, that the estimate will actually equal μ. Hence, it would seem desirable to supplement such a point estimate of μ with some statement as to how close we might reasonably expect the estimate to be. The error, $\bar{X} - \mu$, is the difference between the estimator and the quantity it is supposed to estimate. To examine this error, let us make use of the fact from **Theorem 2** that for large n:

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

is a random variable having approximately the standard normal distribution.

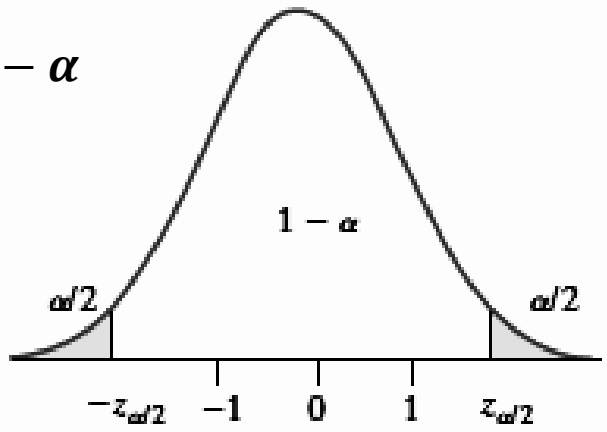
As illustrated in the given figure, for any specified value of α:

$$P\left(-z_{\alpha/2} \leq \frac{(\bar{X} - \mu)}{\sigma/\sqrt{n}} \leq z_{\alpha/2}\right) = 1 - \alpha$$

or, equivalently,

$$P\left(\frac{|\bar{X} - \mu|}{\sigma/\sqrt{n}} \leq z_{\alpha/2}\right) = 1 - \alpha$$

where $z_{\alpha/2}$ is such that the normal curve area to its right equals $\alpha/2$.



The sampling distribution of

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

Now, let E, called the **maximum error of estimate** stand for the maximum of these values of $|\bar{X} - \mu|$. Then, the error $|\bar{X} - \mu|$, will be less than **Maximum error of estimate** E

$$E = z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}$$

with probability $1 - \alpha$.

In other words, if it is intended to estimate μ with the mean of a large ($n \geq 30$) random sample, it can be assert with probability $1-\alpha$ that the error, $|\bar{X} - \mu|$, will be at most $z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}$. The most widely used values for $1-\alpha$ are **0.95 and 0.99.**

Example 3 Specifying a high probability for the maximum error (σ known)

An industrial engineer intends to use the mean of a random sample of size $n = 150$ to estimate the average mechanical skill (as measured by a certain test) of assembly line workers in a large industry. If, on the basis of experience, the engineer can assume that $\sigma = 6.2$ for such data, what can be asserted with probability **0.99** about the maximum size of his error?

Solution:

Substituting $n = 150, \sigma = 6.2$,

α= 1- 0.99=0.01 and then from **Table 3**

At $F(z) = 0.995$ the value of

$z_{\alpha/2} = z_{0.005} = 2.575$

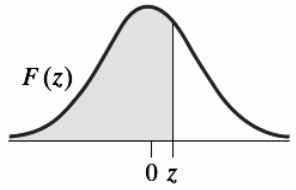
into the the formula

$$E = z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}} = 2.575 \text{ x } \frac{6.2}{\sqrt{150}} = 1.304$$

Thus, the engineer can be asserted with probability 0.99 that his error will be at most 1.30.

Table 3

$$F(z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} e^{-t^2/2}\, dt$$



z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981

The methods discussed so far require that σ be known or that it can be approximated with the sample standard deviation S, thus requiring that n be large. However, if it is reasonable to assume that we are sampling from a normal population, the **Theorem 3** instead of **Theorem 3**, namely on the fact that:

$$t = \frac{\bar{X} - \mu}{S/\sqrt{n}}$$

is a random variable having the t distribution with $\nu = n - 1$ degrees of freedom.

When $\bar{X}$ and S become available, it can be prove with $(1 - \alpha)100\%$ confidence that the error made in using $\bar{x}$ to estimate μ is at most It can be reached the maximum error of estimate, normal population when σ *is* unknown as:

$$E = t_{\alpha/2} \cdot \frac{S}{\sqrt{n}}$$

Example 4 **A 98% confidence bound on the maximum error**

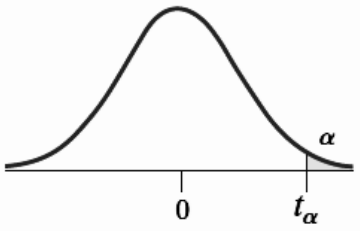
In **six** determinations of the melting point of an aluminum alloy, a chemist obtained a mean of 532.26 °C with a **standard deviation of 1.14 °C**. If this mean is used to estimate the actual melting point of the alloy, what can the chemist confirm with **98%** confidence about the maximum error?

Solution Substituting $\boldsymbol{n = 6,\ S = 1.14}$, and as $\boldsymbol{\alpha = 0.02}$ then from **Table4** $\boldsymbol{t_{0.01} = 3.365}$ and $\boldsymbol{\nu = n - 1 = 5}$ degrees of freedom into the formula for $\boldsymbol{E}$,

$$\boldsymbol{E = t_{\alpha/2}.\frac{S}{\sqrt{n}} = 3.365\ \mathrm{x}\ \frac{1.14}{\sqrt{6}} = 1.57°\mathrm{C}}$$

Thus the chemist can be sure with 98% confidence that his figure for the melting point of the aluminum alloy is off by at most 1.57 °C

Table 4 Values of t_α



ν	$\alpha = 0.10$	$\alpha = 0.05$	$\alpha = 0.025$	$\alpha = 0.01$	$\alpha = 0.00833$	$\alpha = 0.00625$	$\alpha = 0.005$	ν
1	3.078	6.314	12.706	31.821	38.204	50.923	63.657	1
2	1.886	2.920	4.303	6.965	7.650	8.860	9.925	2
3	1.638	2.353	3.182	4.541	4.857	5.392	5.841	3
4	1.533	2.132	2.776	3.747	3.961	4.315	4.604	4
5	1.476	2.015	2.571	3.365	3.534	3.810	4.032	5

Determination of Sample Size

The formula of $E = z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}$ can also be used to determine the sample size that is needed to reach a desired degree of precision. Suppose that the mean of a large random sample is used to estimate the mean of a population, and it is wanted to be sure with probability of $1 - \alpha$ that the error will be at most some prescribed quantity E. As before, if the equation of $E = z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}$ is solved for n then: $n = \left[\frac{z_{\alpha/2} \cdot \sigma}{E}\right]^2$. In order to be able to use this formula $1 - \alpha$, E, and σ must be known and it is often substitute an estimate based on prior data of a similar kind.

Example 5 Selecting the sample size

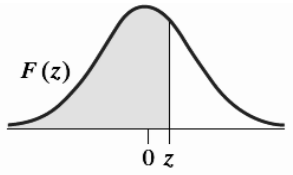
A research worker wants to determine the average time it takes a mechanic to rotate the tires of a car, and he wants to be able to be sure with **95% confidence** that the mean of his sample is off by at most **0.50 minute**. If he can presume from past experience that σ **= 1.6 minutes**, how large a sample will he have to take?

Solution Substituting $E = 0.50, \sigma = 1.6,$ and $z_{0.025} = 1.96$ into the formula for n:

$$n = \left[\frac{z_{\alpha/2} \cdot \sigma}{E}\right]^2 = \left[\frac{1.96 \text{ x } 1.6}{0.5}\right]^2 = 39.33$$

Thus, the research worker will have to time **40** mechanics performing the task of rotating the tires of a car.

$$F(z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} e^{-t^2/2}\, dt$$



z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767

Interval Estimation

It is sometimes preferable to replace point estimates with the so-called **interval estimates**, because point estimates cannot really be expected to match with the quantities that are intended to be estimated.

Referring to the probability statement

$$P\left(-z_{\alpha/2} \leq \frac{(\bar{X}-\mu)}{\sigma/\sqrt{n}} \leq z_{\alpha/2}\right) = 1-\alpha$$

$$P\left(-z_{\alpha/2}\,\frac{\sigma}{\sqrt{n}} \leq (\bar{X}-\mu) \leq z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right) = 1-\alpha$$

$$P\left(\bar{X}-z_{\alpha/2}\,\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X}+z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right) = 1-\alpha$$

This probability statement concerns a random interval covering the unknown parameter μ with probability $1-\alpha$. Accordingly, When the observed value $\bar{x}$ becomes available, then for large sample confidence interval for μ (σ known), the following will be obtained:

$$\bar{X}-z_{\alpha/2}\,\frac{\sigma}{\sqrt{n}} < \mu < \bar{X}+z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

Thus, when a sample has been obtained and the value of $\bar{x}$ has been calculated, it can be claimed with $(1-\alpha)$ 100% confidence that the interval from $\bar{X}-z_{\alpha/2}\,\frac{\sigma}{\sqrt{n}}$ to $\bar{X}+z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$ will be surely contained μ.

Example 5 **Calculating and interpreting a large sample confidence interval**

A random sample of size $\boldsymbol{n = 100}$ is taken from a population with $\boldsymbol{\sigma = 5.1}$. Given that the sample mean is $\boldsymbol{\bar{x} = 21.6}$, construct a **95% confidence** interval for the population mean $\boldsymbol{\mu}$.

Solution Substituting the given values of $\boldsymbol{n, x, \sigma}$, and $\boldsymbol{z_{0.025}}$ for $\boldsymbol{(1-\alpha) = 0.975}$ is $\boldsymbol{= 1.96}$ **(from Table 3)** into the confidence interval formula, it can be obtained:

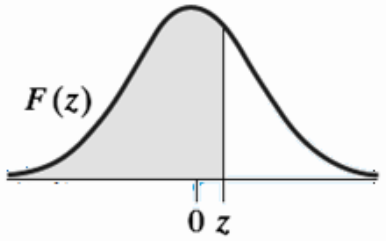
$$\boldsymbol{21.6 - 1.96 \cdot \frac{5.1}{\sqrt{100}} < \mu < 21.6 + 1.96 \cdot \frac{5.1}{\sqrt{100}}}$$

$$\boldsymbol{20.6 < \mu < 22.6}$$

Of course, either the interval from 20.6 to 22.6 contains the population mean $\boldsymbol{\mu}$, or it does not, but we are 95% confident that it does.

Table 3

$$F(z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} e^{-t^2/2}\, dt$$



z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767

The preceding confidence interval formula is exact only for **random samples from normal populations**, but for large samples it will generally provide good approximations. Since $\boldsymbol{\sigma}$ is unknown in most applications, then it will have to make the further approximation of substituting the sample standard deviation $\boldsymbol{S}$ for $\boldsymbol{\sigma}$. Accordingly, the **large sample confidence interval for $\boldsymbol{\mu}$** for samples size $(\boldsymbol{n \geq 30})$, can be as the following:

$$\overline{X} - z_{\alpha/2}\frac{S}{\sqrt{n}} < \mu < \overline{X} + z_{\alpha/2}\frac{S}{\sqrt{n}}$$

For small samples $(\boldsymbol{n < 30})$, small sample confidence interval for $\boldsymbol{\mu}$ of normal population :

$$\overline{X} - t_{\alpha/2}\frac{S}{\sqrt{n}} < \mu < \overline{X} + t_{\alpha/2}\frac{S}{\sqrt{n}}$$

Example 6

The nanopillar height data of $\boldsymbol{n = 50}$, $\boldsymbol{\bar{x} = 305.58\,nm}$, and $\boldsymbol{S^2 = 1,366.86}$ (hence, $\mathbf{S = 36.97\ nm}$), construct ***a 99% confidence interval*** for the population mean of all nanopillars.

Solution Substituting into the confidence interval formula with $\boldsymbol{n = 50}$, $\boldsymbol{\bar{x} = 305.58}$, and $\mathbf{S = 36.97}$, , and $z_{0.005}$ for $\boldsymbol{(1-\alpha) = 0.995}$ is $\boldsymbol{= 2.575}$ **(from Table 3),)** into the confidence interval formula, it can be obtained:

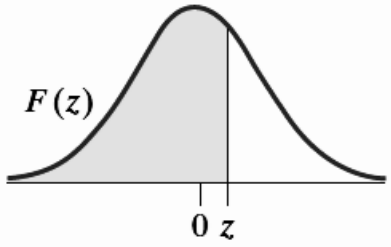
$$\mathbf{305.58 - 2.575 \cdot \frac{36.97}{\sqrt{50}} < \mu < 305.58 + 2.575 \cdot \frac{36.97}{\sqrt{50}}}$$

$$\mathbf{292.12 < \mu < 319.04}$$

It is 99% confident that the interval from 292.12 nm to 319.04 nm contains the true mean nanopillar height.

Table 3

$$F(z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} e^{-t^2/2}\,dt$$



z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986

Example 7 Engineers are making discoveries to create synthetic silk fibers. One research group reports the following statistics for the toughness (MJ/m^3):

$$n = 18, \quad \bar{x} = 22.6 \quad \text{and} \quad S = 15.7$$

Construct **a 95% confidence interval** for the mean toughness of these fibers. Assume that the population is normal distribution.

Solution The sample size is $n = 18$ and $t_{\alpha/2} = t_{0.025}$ from Table 4 at degree of freedom $n - 1 = 17$ is for $n - 1$ =**2.110**.

The 95% confidence formula for μ becomes:

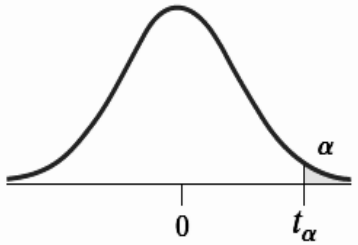
$$\bar{X} - t_{\alpha/2}\frac{S}{\sqrt{n}} < \mu < \bar{X} + t_{\alpha/2}\frac{S}{\sqrt{n}}$$

$$22.6 - 2.11\frac{15.7}{\sqrt{18}} < \mu < 22.6 + 2.11\frac{15.7}{\sqrt{18}}$$

$$14.79 < \mu < 40.41 \text{MJ/m}^3$$

Now, it is 95 % confident that the interval from 14.79 to 36.41 MJ/m3 contains the mean toughness of all possible artificial fibers created

Table 4 Values of t_α



ν	$\alpha = 0.10$	$\alpha = 0.05$	$\alpha = 0.025$	$\alpha = 0.01$	$\alpha = 0.00833$	$\alpha = 0.00625$	$\alpha = 0.005$	ν
16	1.337	1.746	2.120	2.583	2.673	2.813	2.921	16
17	1.333	1.740	2.110	2.567	2.655	2.793	2.898	17
18	1.330	1.734	2.101	2.552	2.639	2.775	2.878	18
19	1.328	1.729	2.093	2.539	2.625	2.759	2.861	19
20	1.325	1.725	2.086	2.528	2.613	2.744	2.845	20