VOLUME VI MARCH, 1908 No. 1

BIOMETRIKA.

THE PROBABLE ERROR OF A MEAN.

BY STUDENT.

Introduction.

ANY experiment may be regarded as forming an individual of a "population" of experiments which might be performed under the same conditions. A series of experiments is a sample drawn from this population.

Now any series of experiments is only of value in so far as it enables us to form a judgment as to the statistical constants of the population to which the experiments belong. In a great number of cases the question finally turns on the value of a mean, either directly, or as the mean difference between the two quantities.

If the number of experiments be very large, we may have precise information as to the value of the mean, but if our sample be small, we have two sources of uncertainty:—(1) owing to the "error of random sampling" the mean of our series of experiments deviates more or less widely from the mean of the population, and (2) the sample is not sufficiently large to determine what is the law of distribution of individuals. It is usual, however, to assume a normal distribution, because, in a very large number of cases, this gives an approximation so close that a small sample will give no real information as to the manner in which the population deviates from normality: since some law of distribution must be assumed it is better to work with a curve whose area and ordinates are tabled, and whose properties are well known. This assumption is accordingly made in the present paper, so that its conclusions are not strictly applicable to populations known not to be normally distributed; yet it appears probable that the deviation from normality must be very extreme to lead to serious error. We are concerned here solely with the first of these two sources of uncertainty.

The usual method of determining the probability that the mean of the population lies within a given distance of the mean of the sample, is to assume a normal distribution about the mean of the sample with a standard deviation equal to $s/\sqrt{n}$, where s is the standard deviation of the sample, and to use the tables of the probability integral.

The Standard Deviation of the Means

• The **standard deviation of the means** (SDOM) is the standard deviation of the means determined from *M* sets of finite data, each consisting of *N* samples of a population.

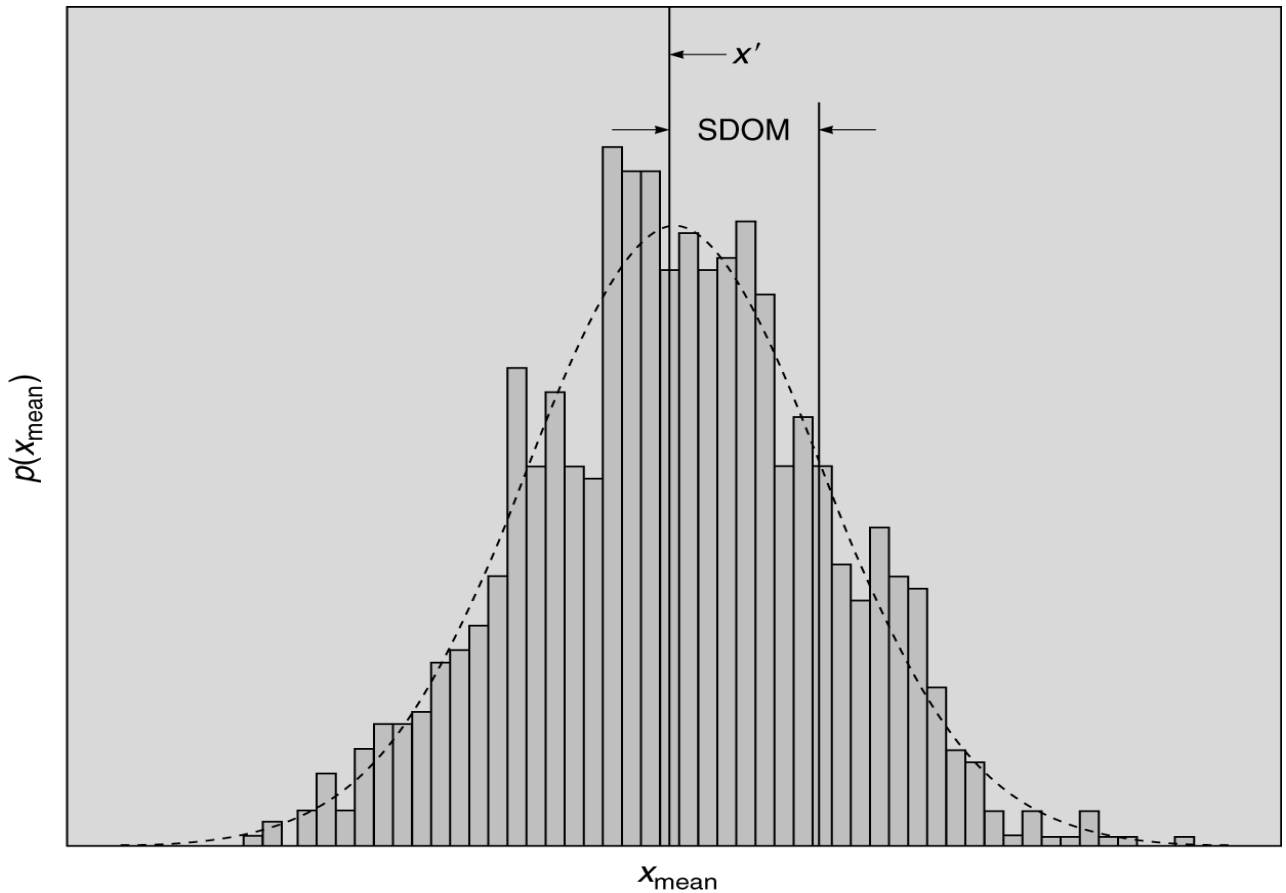


Figure 8.9

SDOM (cont'd)

- The SDOM allows us to estimate x' from $\overline{x}$.

- It can be shown that the SDOM is related to the standard deviation of any one sample by

$$S_{\overline{x}} = S_x \Big/ \sqrt{N}$$

- The SDOM follows a *normal* distribution centered about the mean of the mean values, even if the sampled population is not *normal*.

SDOM (cont'd)

• The SDOM can be used to infer the true mean from the sample mean.

$$x' = \bar{x} \pm t_{\nu,P} S_x / \sqrt{N}$$

• Note that the sample mean approaches the true mean of the population as the sample size, N, becomes very large .

Student's *t* Distribution

- William Gosset, Guinness brewer and statistician, derived Student's *t* distribution, publishing under the pseudonym 'Student' in 1908.
- Student's *t* distribution describes how the members of a *small* sample selected randomly from a normal distribution are distributed.
- There are an infinite number of Student *t* distributions, one for each value of ν, as specified by

$$p(t,\nu) = \frac{\Gamma\left[(\nu+1)/2\right]}{\sqrt{\pi\nu}\,\Gamma(\nu/2)}\left(1+\frac{t^2}{\nu}\right)^{-(\nu+1)/2}$$

Student's *t* and Normal Distributions

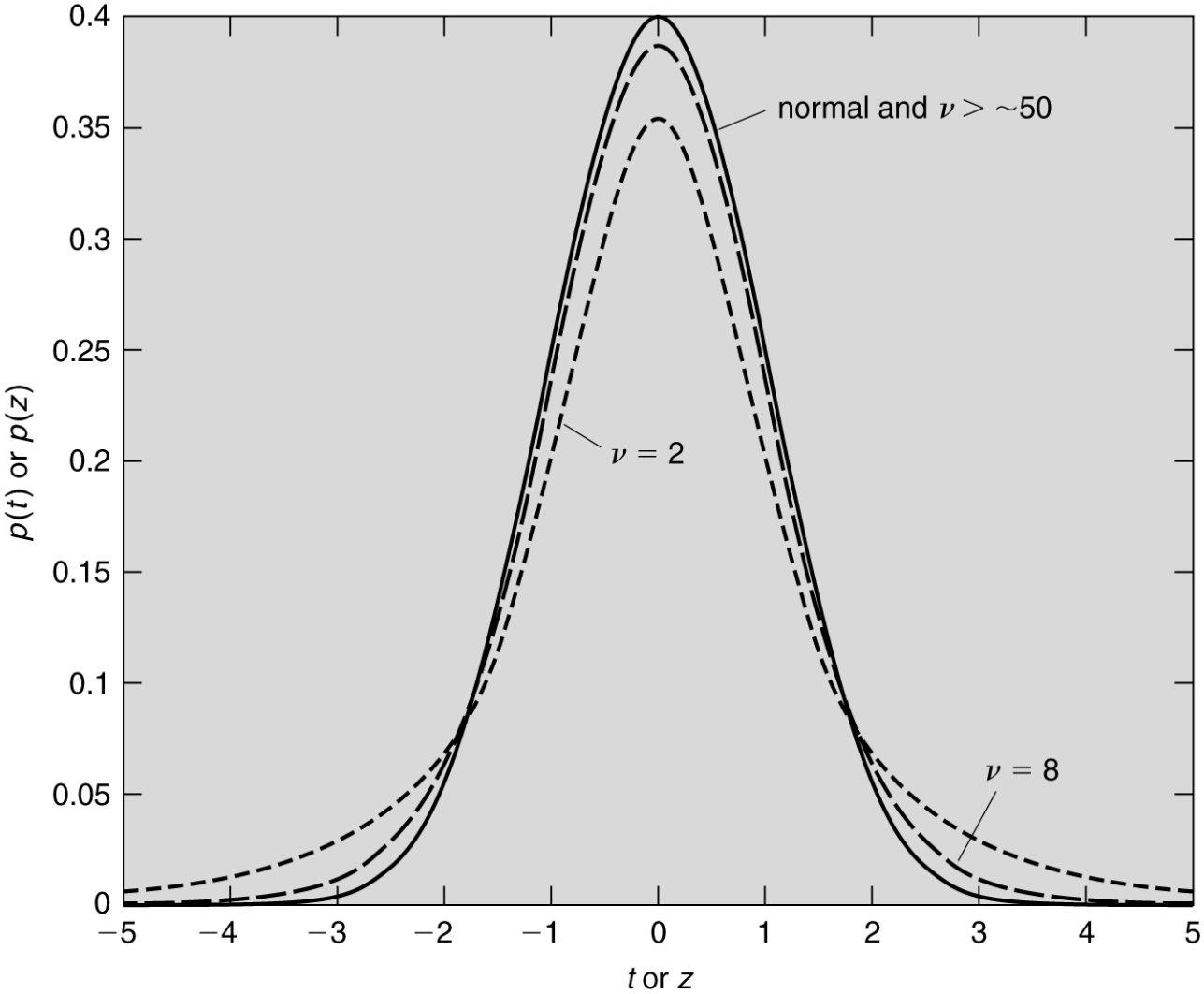


Figure 8.6

t and *z* Comparison

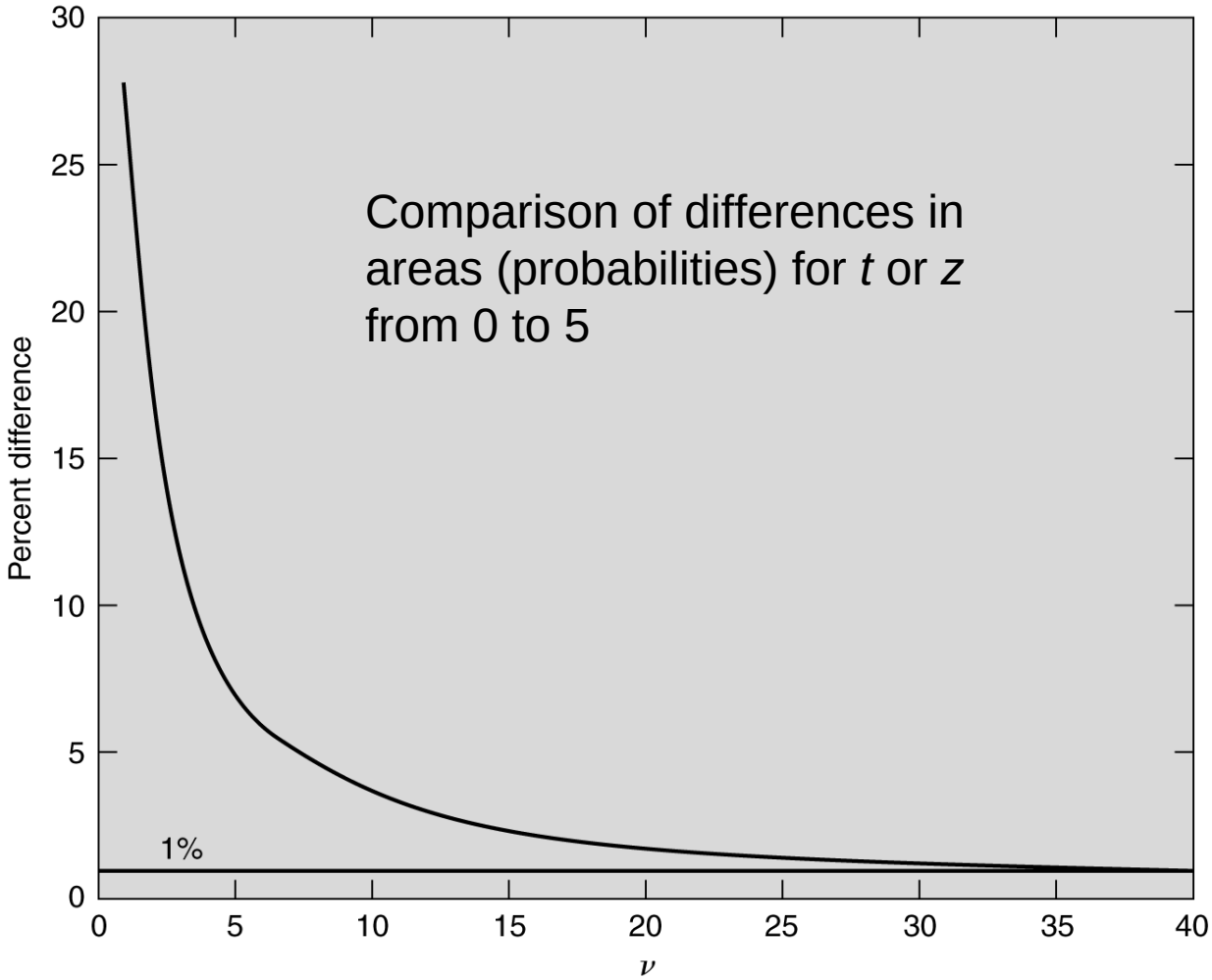


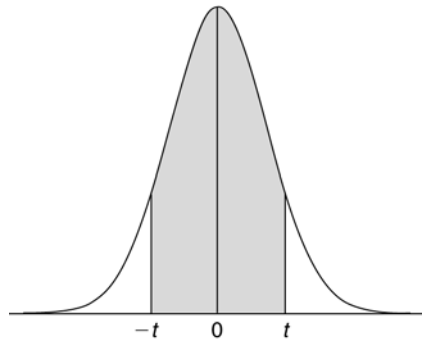
Figure 8.7

Student's *t* Table

Gives the value of *t* for a given ν and P % confidence .

TWO-SIDED STUDENT'S *t* VARIABLE VALUES

ν	$t_{\nu,P=50\,\%}$	$t_{\nu,P=90\,\%}$	$t_{\nu,P=95\,\%}$	$t_{\nu,P=99\,\%}$
1	1.000	6.341	12.706	63.657
2	0.816	2.920	4.303	9.925
3	0.765	2.353	3.192	5.841
4	0.741	2.132	2.770	4.604
5	0.727	2.015	2.571	4.032
6	0.718	1.943	2.447	3.707
7	0.711	1.895	2.365	3.499
8	0.706	1.860	2.306	3.355
9	0.703	1.833	2.262	3.250
10	0.700	1.812	2.228	3.169
11	0.697	1.796	2.201	3.106
12	0.695	1.782	2.179	3.055
13	0.694	1.771	2.160	3.012
14	0.692	1.761	2.145	2.977
15	0.691	1.753	2.131	2.947
16	0.690	1.746	2.120	2.921
17	0.689	1.740	2.110	2.898
18	0.688	1.734	2.101	2.878
19	0.688	1.729	2.093	2.861
20	0.687	1.725	2.086	2.845
21	0.686	1.721	2.080	2.831
30	0.683	1.697	2.042	2.750
40	0.681	1.684	2.021	2.704
50	0.680	1.679	2.010	2.679
60	0.679	1.671	2.000	2.660
120	0.677	1.658	1.980	2.617
∞	0.674	1.645	1.960	2.576



What is *t* for N = 12 ?

Table 8.4

Probabilities for Values of $t_{\nu,P}$

- Sometimes, getting %P from *t* and ν is necessary.

$t_{\nu,P}$	$\%P_{\nu=2}$	$\%P_{\nu=8}$	$\%P_{\nu=100}$
1	57.74	65.34	68.03
2	81.65	91.95	95.18
3	90.45	98.29	99.66
4	94.28	99.61	99.99

Table 8.3

In-Class Example

- What is the probability that a student will score between 75 and 90 on an exam, assuming that the scores are based on 9 students, with a mean of 60 and a standard deviation of 15 ?

Estimates Made Using S_x

- Two different statistical estimates can be made using S_x.

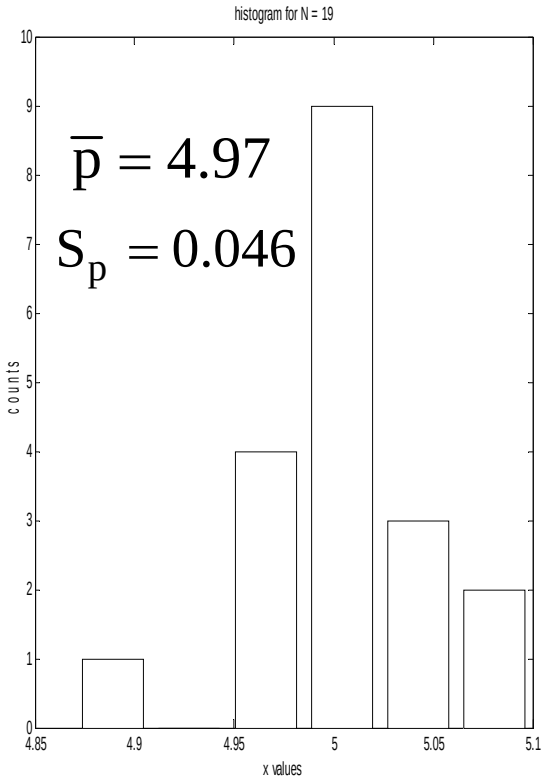
  [1] the value of the next sample value, x_i, where

$$x_i = \bar{x} \pm t_{\nu,P} S_x$$

  [2] the value of the true mean, x‘, where

$$x' = \bar{x} \pm t_{\nu,P} \frac{S_x}{\sqrt{N}}$$

Example 8.5



[1] p range that contains the next measurement with P = 95 %

[2] same but for N = 5

[3] for N = 19 and P = 50 %

[4] p range that contains the true mean

Using MATLAB®

- The command *tpdf(t, ν)* gives the value of p(*t*).
- The command *tcdf(t*, ν)* gives $P(t^*) = \int_{-\infty}^{t^*} p(t)dt$
- The command *tinv(P, ν)* gives *t** from the cdf.
- The command *tinv([1-P]/2, ν)* gives -*t*.
- The command *tinv([1+P]/2, ν)* gives +*t*.