Topic 2. Distributions, hypothesis testing, and sample size determination

2. 1. The Student - t distribution (ST&D pp56 and 77)

Consider a repeated drawing of samples of size n from a normal distribution. For each sample compute $\overline{Y}$, s, $s_{\overline{Y}}$, and another statistic, *t*, where

$$t_{(n-1)} = (\overline{Y} - \mu)/ s_{\overline{Y}}$$

The *t* statistic is the deviation of a normal variable $\overline{Y}$ from its hypothesized mean measured in standard error units. Phrased another way, is the number of standard errors that separate $\overline{Y}$ and μ. The distribution of this statistic is known as the Student's t distribution. There is a unique t distribution for each value of n. For example, for sample size 5, the corresponding t distribution is said to have (5 – 1) degrees of freedom. (df).

Now consider the shape of the frequency distribution of the sampled *t* values. Though symmetric and appearing quite similar to a normal distribution of sample means ($Z = \dfrac{\overline{Y} - \mu}{\sigma_{\overline{Y}}}$), the t distribution will be more variable (i.e. larger dispersion, or broader peak) than Z because $s_{\overline{Y}}$ varies from sample to sample. The larger the sample size, the more precisely $s_{\overline{Y}}$ will estimate $\sigma_{\overline{Y}}$, and the closer *t* approaches Z. *t* values derived from samples of size n ≥ 60 are approximately normally distributed. As n → ∞, t → Z.

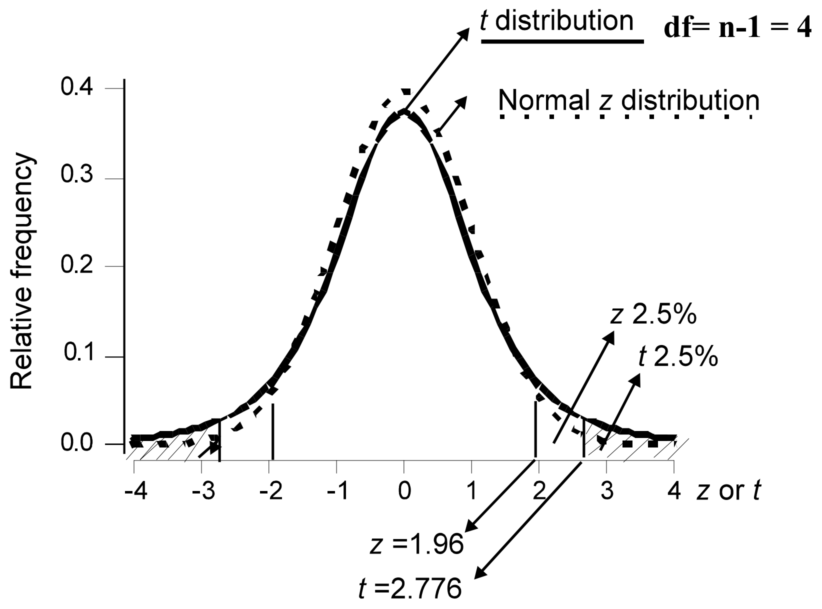


Fig. 1. Distribution of *t* (df=5-1=4) compared to *Z*. The *t* distribution is symmetric and somewhat flatter than *Z*, lying under it at the center and above it in the tails. The increase in the *t* value relative to *Z* is the price we pay for being uncertain about the population variance

2. 2. Confidence limits [S&T p. 77]

Suppose we have a sample {Y_1, ..., Y_n} with mean $\overline{Y}$ drawn from a population with unknown mean μ, and we want to estimate μ. If we manipulate the definition for the *t* statistic we get:

$$\overline{Y} = \mu \pm t_{(n-1)}\ s_{\overline{Y}}$$

Note that μ is a fixed but unknown parameter while $\overline{Y}$ is a known but random statistic. The statistic $\overline{Y}$ is distributed about μ according to the *t* distribution; that is, it satisfies

$$\Pr\{\overline{Y} - t_{\alpha/2,\,n-1}\ s_{\overline{Y}} \le \mu \le \overline{Y} + t_{\alpha/2,\,n-1}\ s_{\overline{Y}}\} = 1 - \alpha$$

Note that for a confidence interval of size α, you must use a *t* value corresponding to an upper percentile of α /2 since both the upper and lower percentiles must be included (see Fig. 1). Therefore the confidence interval is

$$\overline{Y} - t_{\alpha/2,\,n-1}\ s_{\overline{Y}} \le \mu \le \overline{Y} + t_{\alpha/2,\,n-1}\ s_{\overline{Y}}$$

The two terms on either side represent the lower and upper (1- α) **confidence limits** of the mean. The interval between these terms is called the **confidence interval (CI).** For example, in the barley malt extract dataset (see SAS example below), $\overline{Y}$ = 75.94 and $s_{\overline{Y}}$ = 1.227 / √14 = 0.3279. Table A3 gives the $t_{(0.025,\ df=13)}$ value of 2.16, which we multiply by 0.3279 to conclude at the 5% level that μ = 75.94 ± 0.71.

$$\mu = \overline{Y} \pm t_{\frac{0.05}{2},14-1}\ s_{\overline{Y}} = 75.94 \pm 2.16(0.3279) = 75.94 \pm 0.71$$

It is incorrect to say that there is a probability of 95% that the true mean is within 75.94 ± 0.72. If we repeatedly obtained samples of size 14 from the population and constructed these limits for each, we could expect 95% of the intervals to contain the true mean.

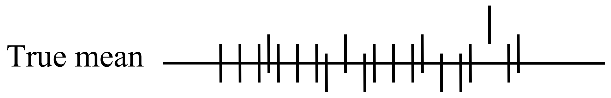


Figure 2. The vertical lines represent 20 95% confidence intervals. One out of 20 intervals does not include the true mean (horizontal line). The confidence level represents the percentage of time the interval covers the true (unknown) parameter value [ST&D p.61]

2. 3. Hypothesis testing [ST&D p. 94]

Another use of the *t* distribution, more in the line of experimental design and ANOVA, is in *hypothesis testing.* Results of experiments are usually not clear-cut and therefore need statistical tests to support decisions between alternative hypotheses. Recall that in this case a *null hypothesis* H_o is established and an alternative H_1 is tested against this. For example, we can use the barley malt data (ST&D p. 30) to test the H_o: $\mu = 78$ against H_1: $\mu \neq 78$.

```
DATA barley;
INPUT extract @@;
CARDS;
77.7 76.0 76.9 74.6 74.7 76.5 74.2 75.4 76.0 76.0 73.9 77.4 76.6 77.3
;
PROC TTEST h0=78 alpha=0.05;
      var extract;
RUN; QUIT;
Statistics

               Lower                Upper
Variable    N  CI mean     Mean    CI mean   Std Dev   Std Err
extract    14  75.234      75.943    76.651   1.2271    0.3279

Variable       DF     t Value     Pr > |t|
extract        13       -6.27       <.0001

Note: PROC UNIVARIATE also produced a t test but only against Ho= 0
```

The formula for ***t*** is as before:

$$t = (\bar{Y} - \mu)/\ s_{\bar{Y}} \qquad \text{With } s_{\bar{Y}} = \sqrt{s^2 / n}$$

For our example: $t = (75.94 - 78) / 0.3279 = -6.27$.

We decide to reject H_o if the probability that we could have obtained this sample from a population with H_o: $\mu = 78$ (satisfying H_o) is less than some pre-assigned number such as **0.05**. This pre-assigned number is known as the *significance level* or **Type I error rate** (**α**). For this example, let's set α = 0.05. From Table A.3, the critical *t* value associated with α = 0.05 (for n =14) is: 2.16.

Since our calculated *t* (-6.27) is larger in absolute value than the critical value $t_{(0.025,13)}$ (2.16) we conclude that the probability of incorrectly rejecting H_o is less than 0.05.

This method is equivalent to calculating a 95% confidence interval around the sample mean ($\bar{Y} \pm t_{0.025,13} * s_{\bar{Y}}$).

In this example, the lower and upper 95% confidence limits for μ under H_o are [75.23 and 76.65]. H_o (78) is not within the confidence interval and we reject $H_{o.}$ A value of $\bar{Y}$ = 75.94 is not expected in a sample of 14 values from a population with H_o $\mu = 78$ and SD: 1.23 unless the particular random sample is a very unusual one.

Errors types

In the previous examples, we defined the value α= 0.05 as the *significance level* of the test as the probability of incorrectly rejecting H_o when it is actually true, a **Type I** error. The other possible error, a **Type II** error, is to incorrectly accept H_o when it is false [S&T p. 118]. The probability of this event is denoted β and is not unique; it depends on the true value of μ. The relationships between hypotheses and decisions can be summarized as follows:

		Null hypothesis	
		Accepted	**Rejected**
Null hypothesis	**True**	Correct decision	**Type I error**
	False	**Type II error**	Correct decision= **Power**

The power of a test (=1- β) is the probability of rejecting the null hypothesis when it is false and the alternative hypothesis is correct. Or in other words a measure of the ability of the test to detect μ_1 (one sample test) **or to detect differences between treatments** (two sample test) **when these differences are real!**

2.3.1. Power of a test for a single sample.

For this discussion, the null hypothesis is

H_0: $\mu_1=\mu_0$ (i.e. the parametric mean of the sampled population is equal to some value μ).

The magnitude of β depends not only on the chosen α but also on how far the alternative parametric mean is from the parametric mean of the null hypothesis. An important concept in connection with hypothesis testing is the **power** of a test (ST&D p.119).

The first equation is for one population with **known σ^2**

$$Power = 1 - \beta = P(Z > Z_{\alpha/2} - \frac{|\mu_1 - \mu_0|}{\sigma_{\bar{Y}}})$$

The 2nd equation is for one population with **unknown σ^2**.

$$Power = 1 - \beta = P(t > t_{\alpha/2} - \frac{|\mu_1 - \mu_0|}{s_{\bar{Y}}})$$

The term $\frac{|\mu_1 - \mu_0|}{s_{\bar{Y}}}$ that is subtracted is the difference between the two means expressed in *standard error units*. By subtracting this term, we are actually moving from the distribution under the null hypothesis to the distribution under the alternative hypothesis (see Figure 3).

As the alternative mean approaches the sample mean, β increases and we lose power (imagine moving the lower curve in figure 3 to the left). This is reasonable since it is more difficult to detect differences that are close than between means that are far apart.

Suppose for example that the true mean is the same as our calculated $\bar{Y}$ = 75.94. What is the power of a test for H_o: μ = 74.88?

Note: an arbitrary alternative mean was selected in this case to obtain an exact β. Usually we do not know the alternative mean, and we just use the difference between means that we would like to detect: e.g. we would like to detect an alternative mean that is 10% higher…

In other words, what is our ability to declare that 74.88 is not the mean of this sampled population if the alternative hypothesis is true?

Fig 3: Since $\alpha = 0.05$, $n = 14$ ($t_{0.025,13} = 2.160$), and $s_{\bar{Y}} = 0.32795$. Using the previous formula for **unknown σ^2**:

$$Power = 1 - \beta = P(t > 2.160 - \frac{|75.94 - 74.88|}{0.32795}) = P(t > -1.072) = 0.85$$

This probability is calculated in Table A3. The probability of the Type II error β we are looking for is the shaded area to the left of the lower curve.

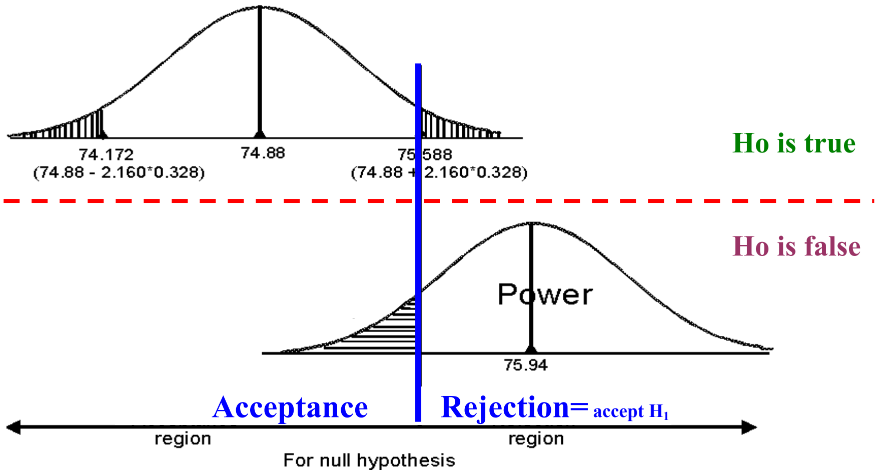


Fig. 3. Type I and Type II errors in the barley data set.

If the alternative hypothesis is true, we have an 85% probability of detecting the difference using a sample of 14 individuals.

2.3.2. Power of the test for the difference between the means of two samples (T-test)

For this discussion, the null hypothesis is H_0: $\mu_1 - \mu_2 = 0$, versus

1) H_1: $\mu_1 - \mu_2 \neq 0$ (two-tailed test), or
2) H_1: $\mu_1 - \mu_2 < 0$ (one-tailed test), or
3) H_1: $\mu_1 - \mu_2 > 0$ (one-tailed test).

[Note: Unless we indicated otherwise, we will focus on the two-tailed test]

The power formula used above **for a single sample with unknown σ^2** *should be modified for the difference* **between two means.**

The *general* power formula for both **equal** and **unequal** sample sizes reads as:

$$Power = P(t > t_{\frac{\alpha}{2}} - \frac{|\mu_1 - \mu_2|}{s_{\bar{Y}1-\bar{Y}2}}) = P(t > t_{\frac{\alpha}{2}} - \frac{|\mu_1 - \mu_2|}{\sqrt{\frac{s^2_{pooled}}{N}}}),$$

where s^2_{pooled} is a weighted average variance: $s^2_{pooled} = \frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{(n_1-1)+(n_2-1)}$

and $N = \frac{n_1 n_2}{n_1 + n_2}$.

Notice in the special case where $\mathbf{n_1 = n_2 = n}$ (equal sample sizes) that the formulas reduce to:

$$s^2_{pooled} = \frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{(n_1-1)+(n_2-1)} = \frac{(n-1)(s_1^2 + s_2^2)}{2(n-1)} = \frac{s_1^2 + s_2^2}{2}$$

$$N = \frac{n_1 n_2}{n_1 + n_2} = \frac{n^2}{2n} = \frac{n}{2}$$

$$Power = P(t > t_{\frac{\alpha}{2}} - \frac{|\mu_1 - \mu_2|}{s_{\bar{Y}1-\bar{Y}2}}) = P(t > t_{\frac{\alpha}{2}} - \frac{|\mu_1 - \mu_2|}{\sqrt{\frac{2s^2_{pooled}}{n}}})$$

Ake note of what this expression indicates: the variance of the difference between two random variables is the sum of their variances (i.e. errors always compound) (ST&D 113-115). If the variances are the same, then the variance of the difference between the two random variables is 2 * average s^2, explaining the multiplication by 2 in the formula of the standard error of the difference between the two means

$$s_{\bar{Y}1-\bar{Y}2} = \sqrt{\frac{2s^2_{pooled}}{n}}$$

The degrees of freedom for the critical t $_{\alpha/2}$ statistic are

General case: $(n_1\text{-}1) + (n_2\text{-}1)$.

For equal sample size: 2*(n-1)

Summary of hypothesis testing [S&T p. 92]

Formulate a meaningful, falsifiable hypothesis for which a test statistic can be computed.
Choose a maximum Type I error rate (α), keeping in mind the power ($1 - \beta$) of the test.
Compute the sample value of the test statistic and find the probability of obtaining, by chance, a value more extreme than that observed.
If the computed, test statistic is greater than the critical, tabular value, reject H_0.

> *Something to consider:*
>
> *H_0 is almost always rejected if the sample size is too large and is almost always not rejected if the sample size is too small.*

2. 4. Determining the sample size

2. 4. 1. Factors affecting replications

There are many factors that affect the decision of how many replications of each treatment should be applied in an experiment. Some of the factors can be put into statistical terms and be determined statistically. Others are non-statistical and depend on experience and knowledge of the subject research matter or available resources for research.

i. ***Nonstatistical factors***: cost and availability of experimental material, the restraints and cost of measurements, the difficulty and applicability of the experimental procedure, the knowledge or experience of the subject matter of research, the objective of the study or intended inference populations.

Example: Use existing information on the experimental units for better sampling selections. Suppose several new varieties are to be tested at different locations. Instead of randomly selecting, say, 10 locations from 100 possible testing sites, one should study these locations first. Maybe it is possible to classify these 100 locations into a few distinct classes of sites based on various physical or environmental attributes of these locations. Thus, it may be appropriate to choose one location randomly from each class to represent certain environmental characteristics for the varietal testing experiment. This example illustrates the point that in order to determine a suitable number of replications, it is not sufficient only to find a number but also the properties of replicated material ought to be considered.

Example: The desired biological response dictates the sample size. In an experiment of testing treatment differences, the important biological difference that needs to be detected must be clearly understood and defined. A relevant sample size can only be determined statistically to assure success with high probability, to detect this significant difference. Thus, in testing an herbicide effect on weed control, what constitutes a meaningful biological effect? Is 60% weed control enough, or do you need 90% weed control?

Between two competitive herbicides, is 5% difference in controlling weeds important enough, or is 10% needed in order be economically relevant?

It should be pointed out that, the smaller the difference to be detected, the larger is the required sample size to obtain the same power. A real difference can always be shown significant statistically, no matter how small it is, by simply increasing sample size (of course, again, if the difference is too small in magnitude, no one will care). But the absence of differences between treatments (Ho) can never be proven with certainty, no matter how large the sample size of the study.

ii. *Statistical factors*: Statistical procedures to determine the number of replications are primarily based on considerations of the required **precision** of an estimator or required **power** of a test. Some commonly used procedures are described in the next section. Note that for a given μ and $s_{\bar{Y}}$, if 2 of the 3 quantities α, β, or the number of observations n are specified then the third one can be determined. In choosing a sample size to detect a particular difference, one must determine the tolerable Type I and Type II errors and then, choose the sample size accordingly.

We will discuss below three basic examples:

1. Sample size estimation for building confidence intervals of certain lengths (2.4.2 and 2.4.3);
2. Sample size estimation for comparing two means, given Type I and II error constraints (2.4.4); and
3. Sample size estimation for building confidence intervals for *standard deviations* using the Chi-square distribution (2.4.5).

2. 4. 2 Sample size for the estimation of a mean with known σ^2 using the *Z* statistic

In this example the objective is parameter estimation rather than on hypothesis testing. The problem is to determine the necessary sample size to estimate a mean by a confidence interval guaranteed to be no longer than a prescribed length.

If the population variance σ^2 is known, or if it is desired to estimate the confidence interval in terms of the population variance, the *Z* statistic for the standard normal distribution may be used. No initial sample is required to estimate the sample size. The sample size can be computed as soon as the required confidence interval length is decided upon. Recall that

$Z = \frac{\bar{Y} - \mu}{\sigma_{\bar{y}}}$ and the confidence interval CI= $\bar{Y} \pm Z_{\alpha/2}\, \sigma_{\bar{Y}}$

Let **d** represent the half-length of the confidence interval

$$d = Z_{\alpha/2}\sigma / \sqrt{n}$$

This can be rearranged to give an expression for n, given α and a desired **d**:

$$n = Z^2_{\frac{\alpha}{2}} \frac{\sigma^2}{d^2}$$

2. 4. 3 Sample size for the estimation of the mean

Stein's Two-Stage Sample [S&T p. 124].

When the variance is unknown and the standard error is used instead of the parametric value, the *t* distribution should replace the normal distribution. The additional complication generated by the use of *t* is that, while the *Z* distribution is independent of the number of replications, the *t* distribution is not. Therefore an iterative approach is required.

Consider a (1 - α)% confidence interval about some mean, μ.

$$\overline{Y} - t_{\alpha/2,\, n-1}\ s_{\overline{Y}} \leq \mu \leq \overline{Y} + t_{\alpha/2,\, n-1}\ s_{\overline{Y}}$$

The length of the (1 - α) % **half-length (d)** of the confidence interval is

$$\mathbf{d} = t_{\alpha/2,\, n-1}\ s_{\overline{Y}} = t_{\alpha/2,\, n-1}\ \ s/\sqrt{n}$$

As in the previous section, this formula can be rearranged to find the sample size **n** required to estimate a mean by a confidence interval no longer than 2d:

$$n = t^2_{\frac{\alpha}{2},n-1} \frac{s^2}{d^2} \approx Z^2_{\frac{\alpha}{2}} \frac{\sigma^2}{d^2}$$ note the similarity between the two formulas

Stein's procedure involves using a pilot study to estimate s^2. Then compute **n** from this formula. Notice that n appears on both sides of this equation.

Before solving this equation via an iterative approach, note that we may also express this equation in terms of the **coefficient of variation**, (CV= s / $\overline{Y}$, ST&D p. 26) and the population mean as:

$$n = t^2_{\frac{\alpha}{2},n-1} \frac{CV^2}{\left(\frac{d}{\overline{Y}}\right)^2}$$

Here **d**/$\overline{Y}$ sets the length of the half confidence interval as a fraction of the estimated mean. For example **d**/$\overline{Y}$ = 0.1 means that the length of **d** (half-length of the confidence interval) should be no larger than one tenth of the population mean.

Example 1 [ST&D, p. 125). The desired maximum length of the 95% confidence interval for the estimation of μ is 10. A preliminary sample of the population yields values of 22, 19, 13, 22, and 23 mm. With $(t_{0.025,\,4})^2 = 2.776^2$=7.71, s^2 = 16.7, and d^2 = 25 we get:

$$n = \frac{16.7 * 7.71}{25} = 5.15$$

Since there is no such thing as a fraction of an observation, we need at least 6 observations.

Now, if the result is much greater than the **n** of the pilot study, **n** will be too far away from the number of degrees of freedom we used to estimate the *t* statistic. In such cases, we must iterate until the equation is satisfied with the same **n** values on both sides of the equal sign.

Example 2: An experimenter wants to estimate the mean height of certain mature plants. From a pilot study of 5 plants, he finds that *s* = 10 cm. What is the required sample size, if he wants to have the total length of a 95% confidence interval about the mean be no longer than 5 cm?

Using $n = t^2_{\frac{\alpha}{2},n-1}\frac{s^2}{d^2}$, the sample size is estimated **iteratively**,

Initial **n**	$t_{\alpha/2=2.5\%,\ df}$	Calculated **n**
5	2.776	$(2.776)^2\,(10)^2/2.5^2 = 123$
123	1.96	$(1.96)^2\,(10)^2 / 2.5^2 = 62$
62	2.00	64
64	2.00	64

Thus, with 64 observations one could estimate the true mean with a precision of 5 cm, at a given α.

Note that if we start with the "*Z*" approximation, then:

$n = Z^2\ \ s^2 / d^2\ = (1.96)^2\,(10)^2 / 2.5^2 = 62$,

which is not too far from the more exact estimate, 64. In fact, one may use the *Z* approximation as a short-cut to bypass the first few rounds of iterations, producing a good estimate of n to then refine using the more appropriate t distribution.

2. 4. 4 Sample size estimation for the comparison of two means

When testing the hypothesis H_o: $\mu_o = \mu_1$, we should take into consideration the possibility of a Type II error and the **power** of a test (1-β). To calculate β we need to know either the alternative mean (μ_1) or the minimum difference we want to detect between the means:

$\delta = |\mu_o\ -\ \mu_1|$.

The appropriate formula for computing n, the required number of observations from **each** treatment, is:

$$n = 2\,(\sigma / \delta)^2\,(Z_{\alpha/2} + Z_\beta)^2$$

For the typical values $\alpha = 0.05$ and $\beta = 0.20$, you find $z_{\alpha/2} = 1.96$, $z_\beta = 0.8416$, and $(Z_{\alpha/2} + Z_\beta)^2 = 7.849$.

A simple way to think this problem is defining δ in terms of σ. For example, we might want to detect a difference between means that are **one σ apart**. In this case:

$n = 2\,(1/1)^2\,(Z_{\alpha/2} + Z_\beta)^2 = 2*7.849$ -> we need ~16 replications per treatment.

To discriminate two means that are **2 σ apart**:

$N = 2 * (½)^2 * 7.849 = 3.92$ ->we need ~ 4 replications per treatment.

Rule of thumb:

Total length CI (one sample)	Separation between means ($\alpha = 0.05$ and $\beta = 0.20$)	Approximate n (per treatment)
2 σ	2 σ	4
1 σ	1 σ	16
½ σ	½ σ	64

The expression above has several obvious difficulties. The first is that we rarely know σ^2 and we cannot honestly use the *Z* statistic. An approximate solution is to multiply the resultant n by a final "correction" factor:

n * (error d.f. +3) / (error d.f. +1) [ST&D p. 123];

where error d.f. = number of pairs -1, for meaningfully paired samples [ST&D p. 106] and 2(n-1) for independent samples. The same equation and correction can be used to estimate the number of blocks in a randomized complete block design analysis of variance [ST&D p. 241]

Another option, in the case of independent samples, is to estimate σ with the pooled sample standard deviation $s_{(pooled)}$ (ST&D p100 Eq.5.8) and replace *Z* by *t*:

$$n = 2\left(\frac{s_{pooled}}{\delta}\right)^2 \left(t_{\frac{\alpha}{2},n1+n2-2} + t_{\beta,n1+n2-2}\right)^2$$

Here **n** is estimated **iteratively**, as in 2.4.3. If **n** is known this equation can be used to estimate the power of the test through the determination of $t_{\beta,\ n1+n2-2}$.

Again, if no estimate of *s* is available, the equation may be expressed in terms of the coefficient of variation, and the difference δ between means as a proportion of the mean:

$$n = 2\,[(\sigma/\mu)\,/\,(\delta/\mu)]^2\,(Z_{\alpha/2} + Z_{\beta})^2 = 2(CV/\delta\%)^2(Z_{\alpha/2} + Z_{\beta})^2$$

Example: Two varieties will be compared for yield, with a previously estimated sample variance of $s^2 = 2.25$. How many replications are needed to detect a difference of 1.5 tons/acre between varieties? Assume $\alpha = 5\%$, and $\beta = 20\%$.

Let's first approximate the required n using the *Z* statistic:

Approximate $\boldsymbol{n} = 2\,(\sigma/\delta)^2\,(Z_{\alpha/2} + Z_{\beta})^2 = 2\,(1.5/1.5)^2(1.96+0.8416)^2 = 15.7$

We'll now use this result as the starting point in an iterative process based on the *t* statistic, where $\boldsymbol{n} = 2\,(s\,/\,\delta)^2\,(t_{\alpha/2} + t_{\beta})^2$ to estimate the sample size iteratively.

Initial n	df = 2n-2	$t_{0.025}$	$t_{0.20}$	estimated n
16	30	2.042	0.854	16.8
17	32	2.037	0.853	16.7

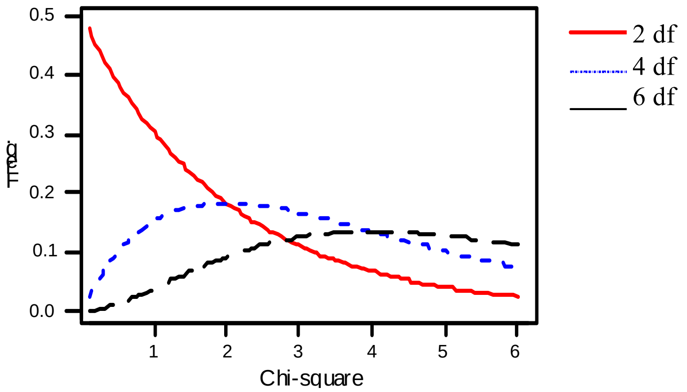
The answer is that there should be 17 replications of each variety. In general, 17 samples per treatment are necessary to discriminate two means that are one standard deviation apart with $\alpha = 5\%$, and $\beta = 20\%$.

2. 4. 5. Sample size to estimate population standard deviation

The Student-t distribution is used to establish confidence intervals around the sample mean as a way to estimating the population mean of a normally distributed random variable. The **chi-squared** distribution is used in a similar way to establish confidence intervals around the sample *variance* as a way to estimate the population variance.

2. 4. 5. 1. The Chi-square distribution (χ^2) [ST&D p. 55]

Figure 4. Distribution of χ^2, for 2, 4, and 6 degrees of freedom.



Relation between Z and χ^2 distributions: The reason that the χ^2 distribution provides a confidence interval for σ^2 is that the Z and χ^2 are related to one another in a simple way: If Z_1, Z_2,... , Z_n are random variables from a *standard* normal distribution then the sum $Z^2_1 + \ldots + Z^2_n$ has a χ^2 distribution with n degrees of freedom.

χ^2 is defined as the sum of squares of independent, normally distributed variables with zero means and unit variances.

$$\boxed{\chi^2_{\alpha,\ \mathrm{df}=1} = Z^2_{(0,1)\ \alpha/2} = \tau^2_{\ \alpha/2,\ \mathrm{df}=\propto}}$$

For example, $\chi^2_{\ 0.05,\ \mathrm{df}=1} = 3.84$ and $Z^2_{(0,1),\ 0.025}$ and $\tau^2_{\ \propto,\ 0.025} = 1.96^2 = 3.84$

Note: Z values from both tails of the N distribution go into the upper tail of the χ^2 for 1 d.f. because of the disappearance of the minus sign in the squaring. For this reason we use α for the χ^2 and $\alpha/2$ for Z and t.

To calculate the formula for the confidence interval for the population standard deviation, we first need to rewrite the sum of squared Z variables:

$$\sum_{i=1}^{n} Z_i^2 = \sum \frac{(Y_i - \mu)^2}{\sigma^2}$$

If we estimate the parametric mean μ with a sample mean, this expression becomes:

$$\sum_{i=1}^{n} Z_i^2 \approx \sum_{i=1}^{n} \frac{(Y_i - \overline{Y})^2}{\sigma^2}$$

Recall now the definition of sample variance:

$$s^2 = \sum_{i=1}^{n} \frac{(Y_i - \overline{Y})^2}{n-1}$$

Therefore,

$$\sum_{i=1}^{n} (Y_i - \overline{Y})^2 = (n-1)s^2$$

And we find that the following statistic has a χ^2_{n-1} distribution:

$$\sum_{i=1}^{n} Z_i^2 \approx \frac{1}{\sigma^2} \sum_{i=1}^{n} (Y_i - \overline{Y})^2 = \frac{(n-1)s^2}{\sigma^2}$$

This expression, which has a χ^2_{n-1} distribution, is important because it provides a **relationship between the sample variance and the parametric variance**.

2. 4. 5. 2. Confidence interval formulas

Based on the expression given above, it follows that a (1-α)% confidence interval for the variance can be derived as follows:

Suppose Y_1, Y_2, ..., Y_n are random variables drawn from a normal distribution with mean μ and variance σ^2. We can make the following probabilistic statement about the ratio $(n-1)\, s^2/\sigma^2$:

$$\boldsymbol{P}\{\ \chi^2_{1-\alpha/2,\ n-1} \le (n-1)\ s^2/\sigma^2\ \le \chi^2_{\alpha/2,\ n-1}\} = 1 - \alpha$$

Simple algebraic manipulation of the quantities in the inequality within brackets yields

$$\boldsymbol{P}\ \{(n-1)\ s^2/\ \chi^2_{\alpha/2,\ n-1} \le \sigma^2 \le (n-1)\ s^2/\ \chi^2_{1-\alpha/2,\ n-1}\} = 1 - \alpha$$

or

$$\boldsymbol{P}\ \{\chi^2_{1-\alpha/2,\ n-1}\ /(n-1) \le s^2/\sigma^2 \le \chi^2_{\alpha/2,\ n-1}/(n-1)\} = 1 - \alpha$$

The 1st form is appropriate when you have an actual estimate of σ^2.
The 2nd form is particularly useful when the precision of s^2 can be expressed in terms of the percent of σ^2.

For example, in the barley malt dataset used before, $s^2 = 1.5057$, and n = 14. If we let α = 0.05 then from the **Table A5** we find $\chi^2_{0.975,13} = 5.01$ and $\chi^2_{0.025,13} = 24.7$. Therefore the 95% confidence interval for σ^2 is [0.79 - 3.91].

<u>Example</u>: What sample size is required if you want to obtain an estimate of σ that you are 90% confident deviates no more than 20% from the true value of σ?

Translating this question into statements of probability:

$\boldsymbol{P}\ (0.8 \le s/\sigma \le 1.2) = .90$ or $\boldsymbol{P}\ (0.64 \le s^2/\sigma^2 \le 1.44) = 0.90$

thus $\chi^2_{(1-\alpha/2,\ n-1)}/(n-1) = 0.64$ or $\chi^2_{(\alpha/2,\ n-1)}/(n-1) = 1.44$

Since χ^2 is not symmetric, the values of n that best satisfy the above two constraints may not be exactly equal if sample size is small. However, small samples generally do not provide good estimate of σ^2 anyway. In practice, then, the required sample size will be large enough to avoid this problem. The actual computation involves an iterative process.

Your starting point is a guess. Just pick an initial n to begin; you will converge to the same solution regardless of where you begin. For this exercise, we've chosen an initial n of 21.

<table>
<tr><th rowspan="2">Size
(n-1)</th><th colspan="2">1 - $\alpha/2$ = 95%</th><th colspan="2">$\alpha/2$ = 5%</th></tr>
<tr><th>$\chi^2_{(n-1)}$</th><th>$\chi^2_{(n-1)}/(n-1)$</th><th>$\chi^2_{(n-1)}$</th><th>$\chi^2_{(n-1)}/(n-1)$</th></tr>
<tr><td>20</td><td>10.90</td><td>0.545</td><td>31.4</td><td>1.57</td></tr>
<tr><td colspan="5">With n=21, we find 0.545 < 0.64 and 1.57 > 1.44. To converge better on the desired values, we need to try a larger n. Let's try n = 41.</td></tr>
<tr><td>40</td><td>26.50</td><td>0.662</td><td>55.8</td><td>1.40</td></tr>
<tr><td colspan="5">Now we have 0.662 > 0.64 and 1.40 < 1.57. We need to go lower. As you continue this process, you will eventually converge on a "best" solution. See the complete table on the next page:</td></tr>
<tr><td><b>35</b></td><td><b>22.46</b></td><td><b>0.642</b></td><td><b>49.8</b></td><td><b>1.42</b></td></tr>
</table>

Thus a rough estimate of the required sample size is approximately 36 (n-1= 35).