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

2.1 The Student - t distribution   [ST&D pp. 56, 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:

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

The *t* statistic is the deviation of a normal variable $\overline{Y}$ from its hypothesized mean measured in standard error units. Phrased another way, it 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 unique value of n. For sample size n, the corresponding t distribution is said to have (n − 1) degrees of freedom (df).

Now consider the shape of the frequency distribution of these sampled *t* values. Though symmetric and appearing quite similar to a normal distribution of sample means ($Z = \dfrac{\overline{Y}_{(n)} - \mu}{\sigma_{\overline{Y}(n)}}$), the t distribution will be more variable (i.e. have a 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. And as n → ∞, t → Z.
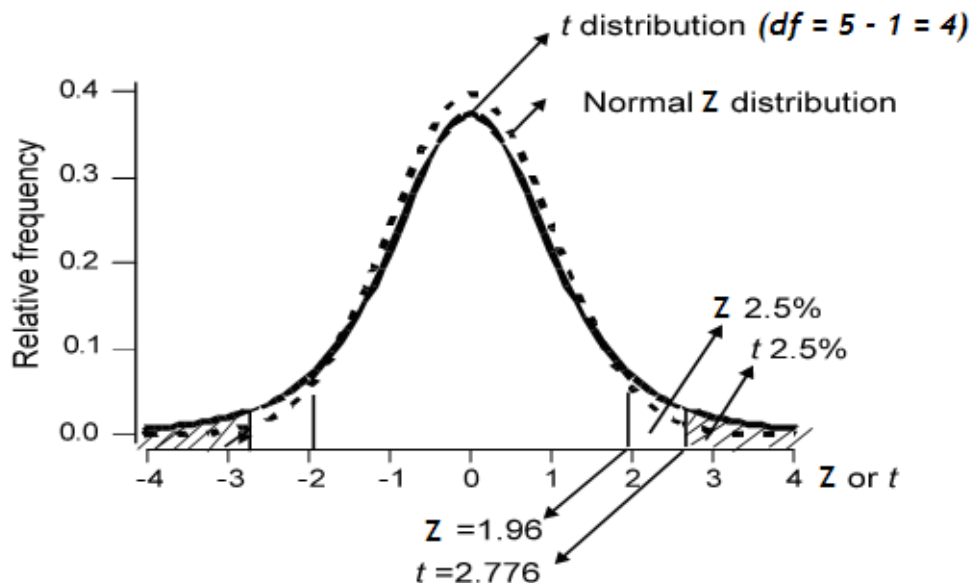

Figure 1 Distribution of *t* (df = 5 − 1 = 4) compared to Z. The t distribution is symmetric and somewhat flatter than the Z, lying under it at the center and above it in the tails. The increase in the t value relative to Z is the price we pay for being uncertain about the population variance

2.2 Confidence limits

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:

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

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

$$P(\overline{Y} - t_{\frac{\alpha}{2}, n-1} s_{\overline{Y}} \leq \mu \leq \overline{Y} + t_{\frac{\alpha}{2}, n-1} s_{\overline{Y}}) = 1 - \alpha$$

Note that for a confidence interval of size α, you must use a t value corresponding to an upper percentile of $\alpha/2$ since both the upper and lower percentiles must be included (see Fig. 1). Consider the term in brackets above:

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

The two terms on either side represent the lower and upper $(1 - \alpha)$ **confidence limits** of the mean. The interval between these terms is called the **confidence interval** (CI). For example, in the barley malt extract data set, $\overline{Y} = 75.94$ and $s_{\overline{Y}} = 1.227 / \sqrt{14} = 0.3279$. Table A.3 gives a corresponding $t_{(0.025, 13)}$ value of 2.16, which we multiply by 0.3279 to conclude at the 5% level that $\mu = 75.94 \pm 0.71$. That is,

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

What does this mean exactly? It is incorrect to say that there is a probability of 95% that the true mean is within the interval 75.94 ± 0.71. The correct way to think about it this: If we repeatedly drew random samples of size n = 14 from the population and constructed a 95% CI for each, we would expect 95% of those intervals (19 out of 20) to contain the true mean.
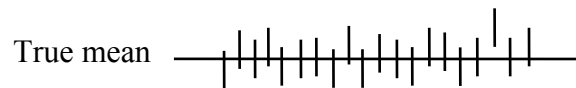

Figure 2 The vertical lines represent 20 95% confidence intervals. One out of 20 intervals does not include the true mean (horizontal line). The confidence level represents the percentage of time the interval covers the true (unknown) parameter value [ST&D p.61].

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_0 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 H_0: $\mu = 78$ against H_1: $\mu \neq 78$.

The definition of *t*, as before:

$$t_{(n-1)} = \frac{\overline{Y}_{(n)} - \mu}{s_{\overline{Y}(n)}}, \text{ with } s_{\overline{Y}} = \frac{s}{\sqrt{n}}$$

So, for our example: $\quad\quad\quad\quad t = (75.94 - 78) / 0.3279 = -6.28$

We reject H_0 if the probability of drawing this sample from a population with $\mu = 78$ (satisfying H_0) is less than some pre-assigned number. This pre-assigned number is known as the significance level or Type I error rate (α). For this example, let's set $\alpha = 0.05$. From Table A.3, the critical t value associated with this significance level (for n =14) is:

$$t_{(\frac{\alpha}{2}, n-1)} = t_{(\frac{0.05}{2}, 14-1)} = 2.160$$

Since our calculated t (- 6.28) is larger, in absolute value, than the critical value $t_{(0.025, 13)}$ (2.160), we conclude that the probability of incorrectly rejecting H_0 is less than 0.05.

This method is equivalent to calculating a 95% confidence interval around the sample mean ($\overline{Y} \pm t_{0.025,13} * s_{\overline{Y}}$). In this example, the lower and upper 95% confidence limits for μ are [75.23, 76.65]. Since 78 (H_0) is not within this confidence interval, we reject H_0. In other words, a value of $\overline{Y} = 75.94$ is not expected for a sample of 14 values from a population with $\mu = 78$ and s = 1.227.

Again, in this example, the value 0.05 is called the *significance level* of the test and it is denoted α. It represents the probability of incorrectly rejecting H_0 when it is actually true, a **Type I** error. The other possible error, a **Type II** error, is to incorrectly accept H_0 when it is false [S&T p. 118]. The probability of this event is denoted β. The relationships between hypotheses, decisions, and errors can be summarized as follows:

H_0	is rejected	is not rejected
is true	**Type I error**	Correct decision
is false	Correct decision	**Type II error**

2.3.1 Power of a test for a single sample.

For this discussion, the null hypothesis is H_0: $\mu_1 = \mu$ (i.e. the parametric mean of the sampled population is equal to some value μ.)

The magnitude of β directly depends not only on the chosen α but also on the actual distance between the two means under consideration in the null hypothesis. As the value of μ_1 approaches that of μ, β increases to a maximum value of $(1 - \alpha)$. An important concept in connection with hypothesis testing is the **power** of a test (ST&D p.119):

$$\text{Power} = 1 - \beta = P(Z > Z_{\frac{\alpha}{2}} - \frac{|\mu_1 - \mu|}{\sigma_{\bar{Y}}}) \text{ OR } P(t > t_{\frac{\alpha}{2}, n-1} - \frac{|\mu_1 - \mu|}{s_{\bar{Y}}})$$

The first equation is for a population with **known σ^2** and the second is for a population with **unknown σ^2** (i.e. s^2 estimates σ^2). The subtracted term is the difference between the two means expressed in units of standard error.

> **Power** is the probability of rejecting H_0 when it is false and the alternative hypothesis is correct. In this case, it is a measure of the ability of the test to differentiate two means.

Suppose, for example, that the true mean of the population is the same as our calculated $\bar{Y} = 75.94$. What is the power of a test for H_0: $\mu = 74.88$? In other words, what is our ability to declare that 74.88 is not the mean of this sampled population? Refer to Figure 3. Since $\alpha = 0.05$ and r = 14, $t_{0.025,13} = 2.160$; and $s_{\bar{Y}} = 0.32795$. Using the previous formula for **unknown σ^2**:

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

The final step in the process above, assigning a value of 0.85 to the probability statement, is done by referencing Table A.3. The Type II error (β) is the shaded area to the left of the lower curve in Figure 3. It is the probability of failing to reject a false H_0.
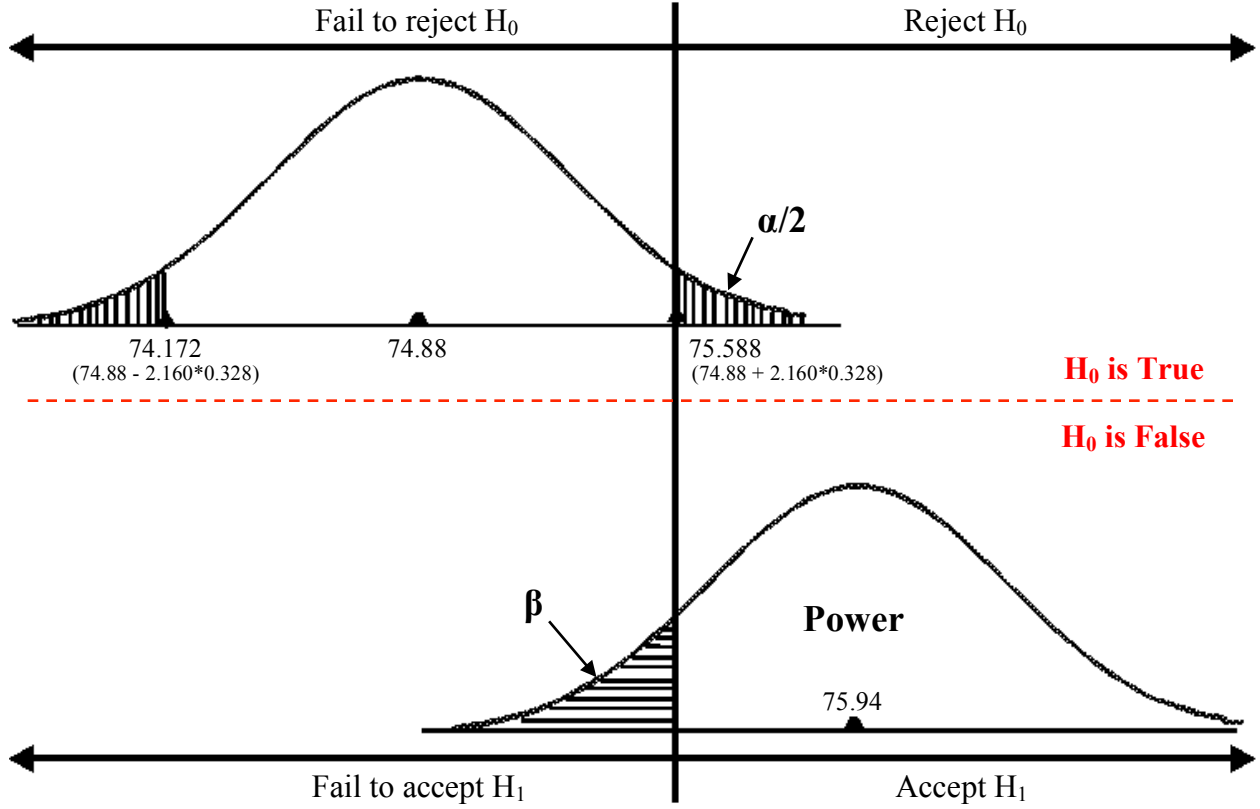
Fig. 3. Type I and Type II errors in the Barley data set.

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), 2) H_1: $\mu_1 - \mu_2 < 0$ (one-tailed test), or 3) H_1: $\mu_1 - \mu_2 > 0$ (one-tailed test). [Note: One-tailed tests are almost never appropriate in basic research, so we will focus in this section on the two-tailed test only.]

The power formula used above **for a single sample with unknown σ^2** *should be modified for the difference* **between two means**. Specifically, the *general* power formula for both **equal** and **unequal** sample sizes is:

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

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

$$\text{and } n_{pooled} = \frac{n_1 n_2}{n_1 + n_2}.$$

Notice in the special case of equal sample sizes (where **n₁ = n₂ = n**), the formulas simplify:

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

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

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

Take 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 derived variable (i.e. the difference of the two random variables) is 2 * average s^2. This is the intuitive explanation for 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: $(n_1-1) + (n_2-1)$
     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 Sample size determination

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 for anyone to care? Between two competitive herbicides, is 5% difference in controlling weeds important enough, or is 10% needed in order to claim superiority of one over the other? It should be pointed out, the smaller the difference to be detected, the larger the required sample size. A positive 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 a lack of difference between treatments (H_0) can never be proven with certainty, no matter how large the sample size in 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 admit the possibility of Type I and Type II errors and choose n accordingly.

We will present 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 the mean with known σ^2, using the Z statistic

Where attention is concentrated on parameter estimation rather than hypothesis testing, a procedure for determining the required number of observations is available for continuous data. 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 true 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:

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

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

$$d = Z_{\frac{\alpha}{2}}\sigma_{\overline{Y}} = Z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}$$

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

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

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 variance, 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 n, the *t* distribution is not. Therefore, an iterative approach is required.

Consider a $(1 - \alpha)\%$ confidence interval about some mean μ:

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

The **half-length (d)** of this confidence interval is therefore:

$$d = t_{\frac{\alpha}{2}, n-1} s_{\bar{Y}} = t_{\frac{\alpha}{2}, n-1} \frac{s}{\sqrt{n}}$$

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:

$$n = t^2_{\frac{\alpha}{2}, n-1} \frac{s^2}{d^2} \approx Z^2_{\frac{\alpha}{2}} \frac{\sigma^2}{d^2}$$

Stein's procedure involves using a pilot study to estimate s^2. Notice that n appears on both sides of this equation. Before solving this equation via an iterative approach, note that we may also express it in terms of the **coefficient of variation** $(CV = s / \bar{Y}$, ST&D p. 26):

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

Here $(d / \bar{Y})$ sets the length of the confidence interval as a fraction of the estimated mean. For example $(d / \bar{Y}) = 0.1$ 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 mm. 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:

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

Therefore, we need slightly more than five observations. Since there is no such thing as a fraction of an observation, we conclude that we need six observations.

Now, if the result is much greater than the n of the pilot study, we have a problem because the n's on the two sides of the equation will not agree. In other words, n will be too far away from the number of degrees of freedom used to determine 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 $n = t^2_{\frac{\alpha}{2}, n-1} \dfrac{s^2}{d^2}$, the sample size is estimated **iteratively**:

Initial n	$t_{0.025, n-1}$	Calculated n
5	2.776	$(2.776)^2 (10)^2 / 2.5^2 = 123.3$
124	1.96	$(1.96)^2 (10)^2 / 2.5^2 = 61.5$
62	2.00	64
64	2.00	64

Thus, with 64 observations, one could estimate the true mean with a precision 5 cm, at the given α. Note that if we started with a 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_0: $\mu_1 = \mu_2$, we should also take into account the possibility of a Type II error and, thereby, the power of the test $(1 - \beta)$. To calculate β, we need to know either the alternative mean or at least the minimum difference we wish to detect between the means $(\delta = |\mu_1 - \mu_2|$. 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 $(Z_{\alpha/2} + Z_\beta)^2 = 7.849$.

So, to discriminate two means that are 2σ apart, approximately 4 replications per treatment are required $(2 * (\frac{1}{2})^2 * 7.849 = 3.92)$. If the two means are 1 or 0.5 standard deviations apart, the required n increases to 16 and 63, respectively.

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:

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

The problem may also be obviated by defining δ in terms of σ. For example, we might want to detect a difference between means of one standard deviation in size. In this case, $n = 2\ (Z_{\alpha/2} + Z_{\beta})^2$.

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 α = 5% and β = 20%.

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

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

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

Initial n	df = 2n-2	$t_{0.025,2n-2}$	$t_{0.20,2n-2}$	Calculated 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 α = 5% and β = 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 of 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 of estimating the population variance.

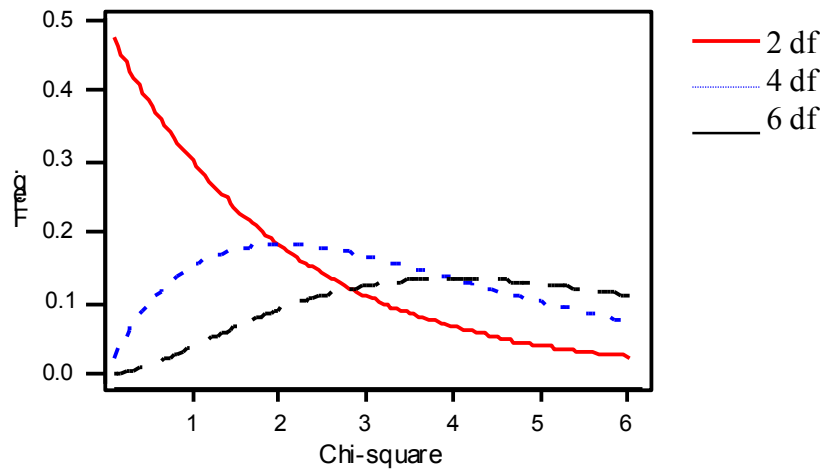### 2.4.5.1 The Chi- square distribution [ST&D p. 55]

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

Relation between the normal and chi-square distributions: The reason that the chi square distribution provides a confidence interval for σ^2 is that 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 $Z_1^2 + Z_2^2 + ... + Z_n^2$ 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.

Therefore:

$$\chi^2_{\alpha, df=1} = Z^2_{\frac{\alpha}{2}} = t^2_{\frac{\alpha}{2}, df=\infty}$$

$$\text{e.g. } \chi^2_{0.05,1} = 3.84, \ Z^2_{\frac{\alpha}{2}} = 1.96^2 = 3.84, \text{ and } t^2_{\frac{\alpha}{2}, df=\infty} = 1.96^2 = 3.84$$

Note that 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.

Let's rewrite this sum of squared Z variables:

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

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

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

Recall now the definition of sample variance:

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

Therefore,

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

And we find:

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

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 $(1 - \alpha)\%$ 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 $(n-1)\,s^2/\sigma^2$:

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

Simple algebraic manipulation of the quantities within the brackets yields

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

OR

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

The first form is particularly useful when the desired precision of s^2 can be expressed in terms of a proportion of σ^2. The second form is appropriate when you have an actual estimate of σ^2. For example, in the barley data from before, $s^2 = 1.5057$ and $n = 14$. If we let $\alpha = 0.05$, then from Table A.5 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].

Example: 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:

$$P\,(0.8 \le s / \sigma \le 1.2) = 0.90 \quad OR \quad P\,(0.64 \le s^2 / \sigma^2 \le 1.44) = 0.90$$

thus

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

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 excerise, we've chosen an initial n of 21.

n	df (n-1)	$1 - \alpha/2 = 95\%$		$\alpha/2 = 5\%$	
		$\chi^2_{(n-1)}$	$\chi^2_{(n-1)}/(n-1)$	$\chi^2_{(n-1)}$	$\chi^2_{(n-1)}/(n-1)$
21	20	10.90	**0.545**	31.4	**1.57**

With this n, 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.

n	df (n-1)	$1 - \alpha/2 = 95\%$		$\alpha/2 = 5\%$	
		$\chi^2_{(n-1)}$	$\chi^2_{(n-1)}/(n-1)$	$\chi^2_{(n-1)}$	$\chi^2_{(n-1)}/(n-1)$
41	40	26.50	0.662	55.8	1.40

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:

n	df (n-1)	$1 - \alpha/2 = 95\%$		$\alpha/2 = 5\%$	
		$\chi^2_{(n-1)}$	$\chi^2_{(n-1)}/(n-1)$	$\chi^2_{(n-1)}$	$\chi^2_{(n-1)}/(n-1)$
21	20	10.90	0.545	31.4	1.57
41	40	26.50	0.662	55.8	1.40
31	30	18.50	0.616	43.8	1.46
36	**35**	**22.46**	**0.642**	**49.8**	**1.42**
35	**34**	**21.66**	**0.637**	**48.6**	**1.43**

Thus a rough estimate of the required sample size is approximately 35.