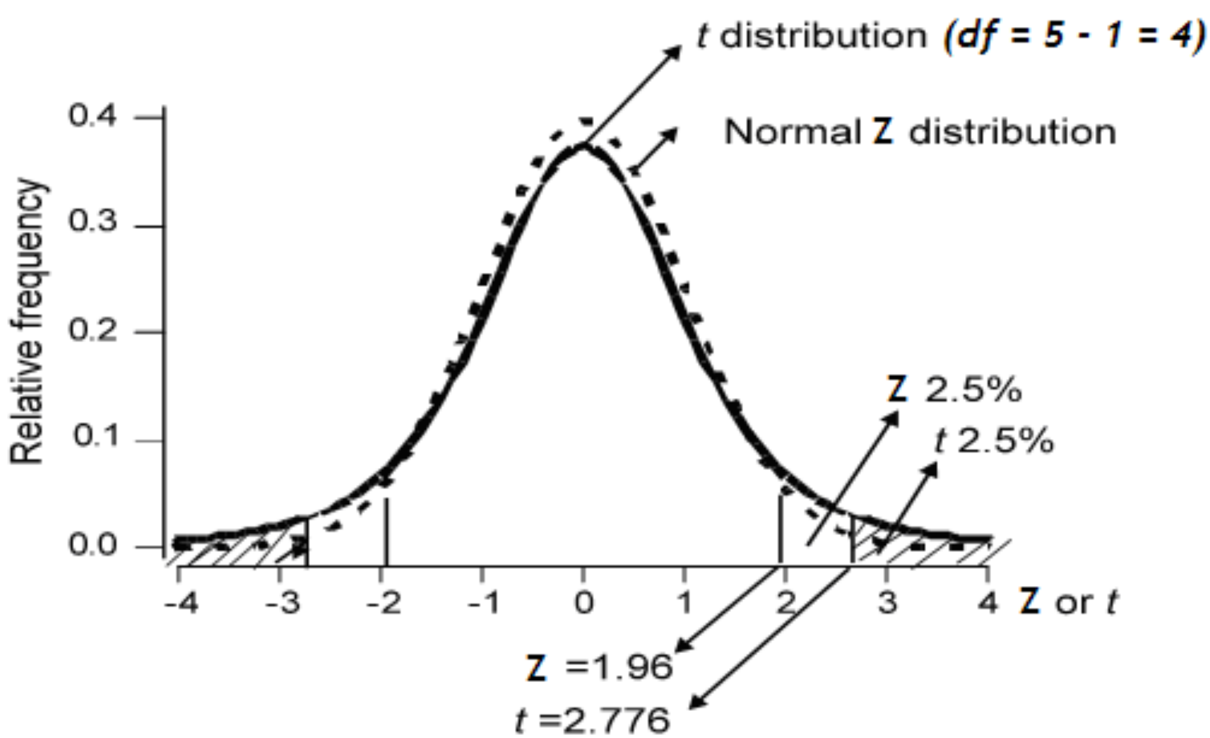
Lecture 3
Topic 2: Distributions, hypothesis testing, and sample size determination

The Student - t distribution

Consider a repeated drawing of samples of size n from a normal distribution of mean μ. 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.



For any given value of α, $|t_{crit}|$ is always larger than $|Z_{crit}|$. This is the price we pay for being uncertain about the population variance

Confidence limits based on sample statistics

Taking into account the imperfect information provided by sampling, the estimated value of any population parameter (λ) takes the general form:

$$\boxed{\lambda = (\text{Estimated } \lambda) \pm (\text{Critical Value} * \text{Standard error of the estimated } \lambda)}$$

So, for a population mean estimated via a sample mean:

$$\mu = \overline{Y} \pm t_{\frac{\alpha}{2},n-1} * s_{\overline{Y}}$$

The statistic $\overline{Y}$ is distributed about μ according to the *t* distribution, satisfying:

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

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).

$$(1 - \alpha) \text{ CI for } \mu = [\overline{Y} - t_{\frac{\alpha}{2},n-1} s_{\overline{Y}}\ ,\ \overline{Y} + t_{\frac{\alpha}{2},n-1} s_{\overline{Y}}]$$

Example: Data set of 14 barley malt extract values ($\overline{Y} = 75.94$, $s_{\overline{Y}} = 1.227 / \sqrt{14} = 0.3279$).

By t-table or R: $t_{\frac{\alpha}{2},n-1} = t_{0.025,13} = 2.16$

95% CI for $\mu = 75.94 \pm 2.16(0.3279) = 75.94 \pm 0.71 = [75.23, 76.65]$

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.

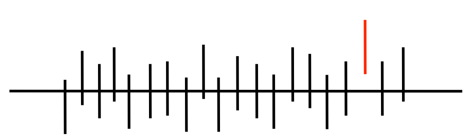
True mean 

TABLE A.3
Values of t

	Probability of a numerically larger value of t								
df	**.5**	**.4**	**.3**	**.2**	**.1**	**.05**	**.02**	**.01**	**.001**
1	1.000	1.376	1.963	3.078	6.314	12.706	31.821	63.657	636.619
2	0.816	1.061	1.386	1.886	2.920	4.303	6.965	9.925	31.598
3	0.765	0.978	1.250	1.638	2.353	3.182	4.541	5.841	12.941
4	0.741	0.941	1.190	1.533	2.132	2.776	3.747	4.604	8.610
5	0.727	0.920	1.156	1.476	2.015	2.571	3.365	4.032	6.859
6	0.718	0.906	1.134	1.440	1.943	2.447	3.143	3.707	5.959
7	0.711	0.896	1.119	1.415	1.895	2.365	2.998	3.499	5.405
8	0.706	0.889	1.108	1.397	1.860	2.306	2.896	3.355	5.041
9	0.703	0.883	1.100	1.383	1.833	2.262	2.821	3.250	4.781
10	0.700	0.879	1.093	1.372	1.812	2.228	2.764	3.169	4.587
11	0.697	0.876	1.088	1.363	1.796	2.201	2.718	3.106	4.437
12	0.695	0.873	1.083	1.356	1.782	2.179	2.681	3.055	4.318
13	0.694	0.870	1.079	1.350	1.771	2.160	2.650	3.012	4.221
14	0.692	0.868	1.076	1.345	1.761	2.145	2.624	2.977	4.140
15	0.691	0.866	1.074	1.341	1.753	2.131	2.602	2.947	4.073
16	0.690	0.865	1.071	1.337	1.746	2.120	2.583	2.921	4.015
17	0.689	0.863	1.069	1.333	1.740	2.110	2.567	2.898	3.965
18	0.688	0.862	1.067	1.330	1.734	2.101	2.552	2.878	3.922
19	0.688	0.861	1.066	1.328	1.729	2.093	2.539	2.861	3.883
20	0.687	0.860	1.064	1.325	1.725	2.086	2.528	2.845	3.850
21	0.686	0.859	1.063	1.323	1.721	2.080	2.518	2.831	3.819
22	0.686	0.858	1.061	1.321	1.717	2.074	2.508	2.819	3.792
23	0.685	0.858	1.060	1.319	1.714	2.069	2.500	2.807	3.767
24	0.685	0.857	1.059	1.318	1.711	2.064	2.492	2.797	3.745
25	0.684	0.856	1.058	1.316	1.708	2.060	2.485	2.787	3.725
26	0.684	0.856	1.058	1.315	1.706	2.056	2.479	2.779	3.707
27	0.684	0.855	1.057	1.314	1.703	2.052	2.473	2.771	3.690
28	0.683	0.855	1.056	1.313	1.701	2.048	2.467	2.763	3.674
29	0.683	0.854	1.055	1.311	1.699	2.045	2.462	2.756	3.659
30	0.683	0.854	1.055	1.310	1.697	2.042	2.457	2.750	3.646
40	0.681	0.851	1.050	1.303	1.684	2.021	2.423	2.704	3.551
60	0.679	0.848	1.046	1.296	1.671	2.000	2.390	2.660	3.460
120	0.677	0.845	1.041	1.289	1.658	1.980	2.358	2.617	3.373
∞	0.674	0.842	1.036	1.282	1.645	1.960	2.326	2.576	3.291
df	**.25**	**.2**	**.15**	**.1**	**.05**	**.025**	**.01**	**.005**	**.0005**
	Probability of a larger positive value of t								

Source: This table is abridged from Table III of Fisher and Yates, *Statistical Tables for Biological, Agricultural, and Medical Research,* published by Oliver and Boyd Ltd., Edinburgh, 1949, by permission of the authors and publishers.

Hypothesis testing and power

Example: Data set of 14 barley malt extract values ($\overline{Y} = 75.94$, $s_{\overline{Y}} = 1.227 / \sqrt{14} = 0.3279$).

1. Choose a null hypothesis: Test H_0: $\mu = 78$ versus H_1: $\mu \neq 78$.
2. Choose a significance level: Assign $\alpha = 0.05$.
3. Calculate the test statistic:

$$t = \frac{\overline{Y} - \mu}{s_{\overline{Y}}} = \frac{75.94 - 78.00}{0.3279} = -6.28$$

4. Compare the absolute value of the test statistic to the critical statistic:

$$|-6.28| > 2.16$$

5. Since the absolute value of the test statistic is larger, we reject H_0.

This is equivalent to calculating a 95% confidence interval around $\overline{Y}$.

Since 78 (H_0) is not within the 95% CI [75.23, 76.65], we reject H_0.

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

α = significance level = Type I error rate
= the probability of incorrectly rejecting a true H_0

β = Type II error rate
= the probability of failing to reject a false H_0

Power = $(1 - \beta)$
= the probability of correctly rejecting a false H_0

Power of a test for a single sample

$$\text{Power} = 1 - \beta = P(Z > Z_{\frac{\alpha}{2}} - \frac{|\mu_1 - \mu_0|}{\sigma_{\bar{Y}}}) \text{ OR } P(t > t_{\frac{\alpha}{2}, n-1} - \frac{|\mu_1 - \mu_0|}{s_{\bar{Y}}})$$

Example: Using the same barley data, what is the power of a test for H_0: $\mu = 74.88$? Again, $\alpha = 0.05$, r = 14, $t_{0.025,13} = 2.160$, and $s_{\bar{Y}} = 0.32795$.

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

The magnitude of β depends upon:

1. The Type I error rate (α)
2. The actual distance between the two means under consideration
3. The number of observations (n) → $s_{\bar{Y}} = \frac{s}{\sqrt{n}}$

For a given s and detection distance, if any two of the quantities α, β, or n are specified, the third is determined.

Use sufficient replication to keep Type I and Type II errors under their desired limits.

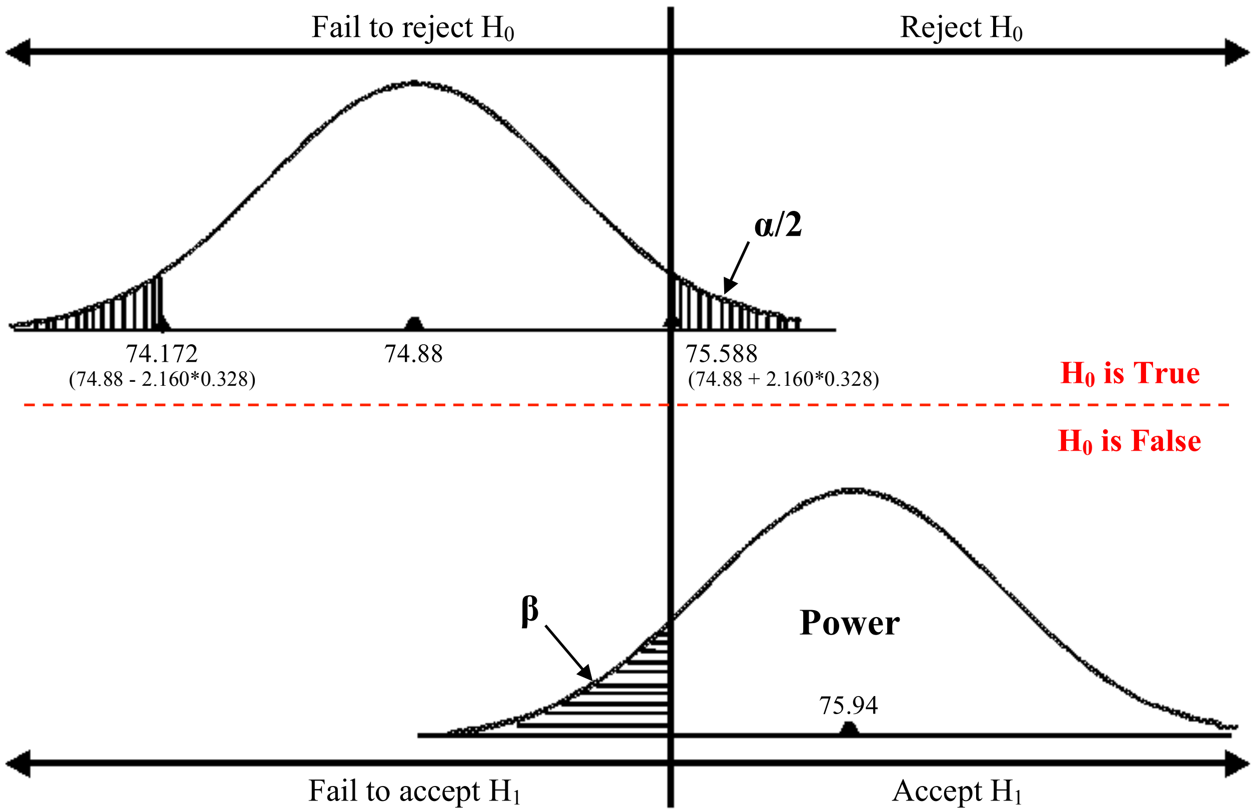
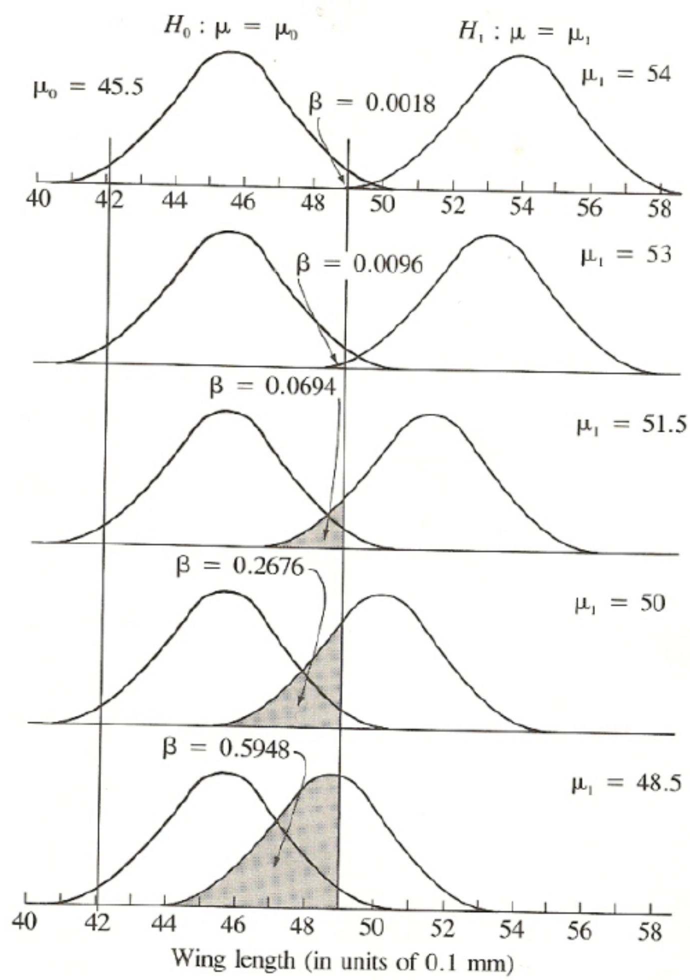


Fig. 3. Type I and Type II errors in the Barley data set.

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.

Power curves

Variation of power as a function of the distance between the alternative hypotheses (Biometry Sokal and Rohlf)



Power of the test for the difference between the means of two samples

H_0: $\mu_1 - \mu_2 = 0$, versus: 1) H_1: $\mu_1 - \mu_2 \neq 0$ (two-tailed test)
2) H_1: $\mu_1 - \mu_2 < 0$ or H_1: $\mu_1 - \mu_2 > 0$ (one-tailed tests)

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

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

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

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

In the special case of equal sample sizes (where $\mathbf{n_1 = n_2 = n}$), the formulas simplify:

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

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

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

The variance of the difference between two random variables is the sum of their variances (i.e. errors always compound).

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)

Sample size for estimating μ, when σ^2 is known (using the Z statistic)

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 may be used.

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

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

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

This can be rearranged to give an expression for n:

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

For standard α = 0.05, $n = Z^2_{\frac{\alpha}{2}} \frac{\sigma^2}{d^2} = Z^2_{\frac{0.05}{2}} \frac{\sigma^2}{d^2} = 1.96^2 \frac{\sigma^2}{d^2} = 3.84 \frac{\sigma^2}{d^2}$

If $d = \sigma$, n ≈ 4   If $d = 0.5\sigma$, n ≈ 16   If $d = 0.25\sigma$, n ≈ 64

This equation can be re-expressed in terms of the coefficient of variation:

$$n = Z^2_{\frac{\alpha}{2}} \frac{\left(\frac{\sigma}{\mu}\right)^2}{\left(\frac{d}{\mu}\right)^2} = Z^2_{\frac{\alpha}{2}} \frac{CV^2}{\left(\frac{d}{\mu}\right)^2}$$

Example: The CVs of yield trials at our experimental station are never greater than 15%. How many replications are needed to construct a 95% CI for the true mean with a total length of no more than 10% of the true mean?

2d = 0.10, so d = 0.05 $n = 1.96^2\,(0.15^2 / 0.05^2) = 34.6 \approx 35$

Sample size for estimating μ, when σ^2 is unknown

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

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

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

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

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

Stein's Two-Stage procedure involves using a pilot study to estimate s^2.

Example: An breeder wants to estimate the mean height of certain mature plants. From a pilot study of 5 plants, she finds that s = 10 cm. What is the required sample size, if she 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	$t_{0.025,n-1}$	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 of 5 cm, at the given α. Note that if we started with a Z approximation, then:

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

Sample size estimation for the comparison of two means

When testing the hypothesis H_0: $\mu_1 = \mu_2$, we can take into account the possibilities of Type I and Type II errors *simultaneously*.

To calculate n, 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 $\alpha = 0.05$ and $\beta = 0.20$: $Z_{\alpha/2} = 1.96$, $Z_\beta = 0.8416$, and $(Z_{\alpha/2} + Z_\beta)^2 = 7.849 \approx 8$

If $\delta = 2\sigma$, $n \approx 4$
If $\delta = 1\sigma$, $n \approx 16$
If $\delta = 0.5\sigma$, $n \approx 64$

We rarely know σ^2 and must estimate it via sample variances:

$$n = 2\left(\frac{s_{pooled}}{\delta}\right)^2 \left(t_{\frac{\alpha}{2},n1+n2-2} + t_{\beta,n1+n2-2}\right)^2, \text{ where } s_{pooled} = \sqrt{\frac{s_1^2 + s_2^2}{2}}$$

Here, n is estimated **iteratively**. If no estimate of *s* is available, the equation may be expressed in terms of the CV and the difference δ as a proportion of the mean:

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

We can also **define δ in terms of σ**.

Example: Two varieties are 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 $\beta = 20\%$.

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

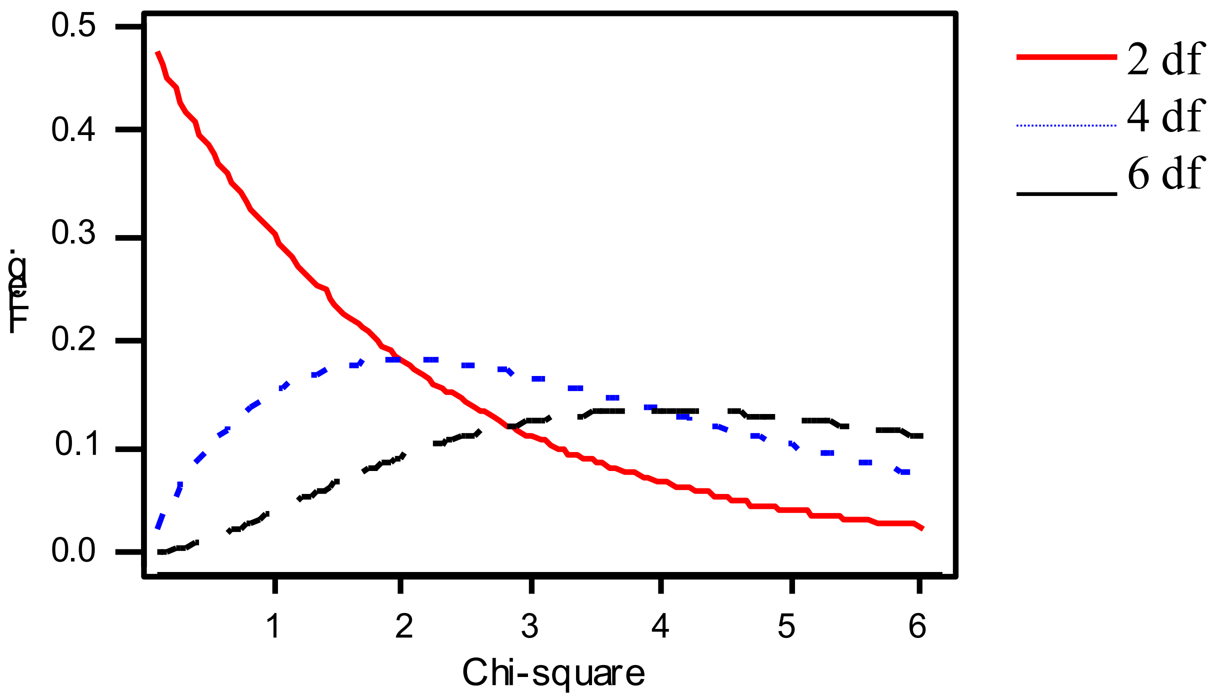
Initial n	df = 2n-2	$t_{0.025,2n-2}$	$t_{0.20,2n-2}$	Calculated n
16	30	2.042	0.854	16.8
17	32	2.037	0.853	16.7

The answer is that there should be 17 replications of each variety.

<u>For the interested</u>: Sample size to estimate population standard deviation

The chi-squared (χ^2) distribution is used to establish confidence intervals around the sample variance as a way of estimating the true, unknown population variance.

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



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

$$\chi^2_{df=n} \equiv \sum_{i=1}^{n} Z_i^2$$

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

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

Resuming…

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

If we estimate the parametric mean μ with a sample mean, we obtain:

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

…due to: $$s^2 = \sum_{i=1}^{n} \frac{(Y_i - \overline{Y})^2}{n-1} \quad \rightarrow \quad \sum_{i=1}^{n} (Y_i - \overline{Y})^2 = (n-1)s^2$$

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

Confidence interval for σ^2

We can make the following probabilistic statement about the ratio (n-1) s^2/σ^2:

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

Simple algebraic manipulation of the quantities within the brackets yields

$$P\left(\frac{\chi^2_{1-\frac{\alpha}{2},n-1}}{(n-1)} \le \frac{s^2}{\sigma^2} \le \frac{\chi^2_{\frac{\alpha}{2},n-1}}{(n-1)} \right) = 1 - \alpha \quad OR \quad P\left(\frac{(n-1)s^2}{\chi^2_{\frac{\alpha}{2},n-1}} \le \sigma^2 \le \frac{(n-1)s^2}{\chi^2_{1-\frac{\alpha}{2},n-1}} \right) = 1 - \alpha$$

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:

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

thus

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

n	df (n-1)	1 - α/2 = 95%		α/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.