

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

2. 1. The Student - t distribution (read ST&D pg 56 and 77)

Consider a repeated drawing of samples of size $r = 5$ from a normal distribution. For each sample compute \bar{Y} , s , $s_{\bar{Y}}$, and another statistic, t , where

$$t = (\bar{Y} - \mu) / s_{\bar{Y}} \quad (\text{Remember } z = (\bar{Y} - \mu) / \sigma_{\bar{Y}})$$

The t statistics is the deviation of a normal variable \bar{Y} from its hypothesized mean measured in standard error units

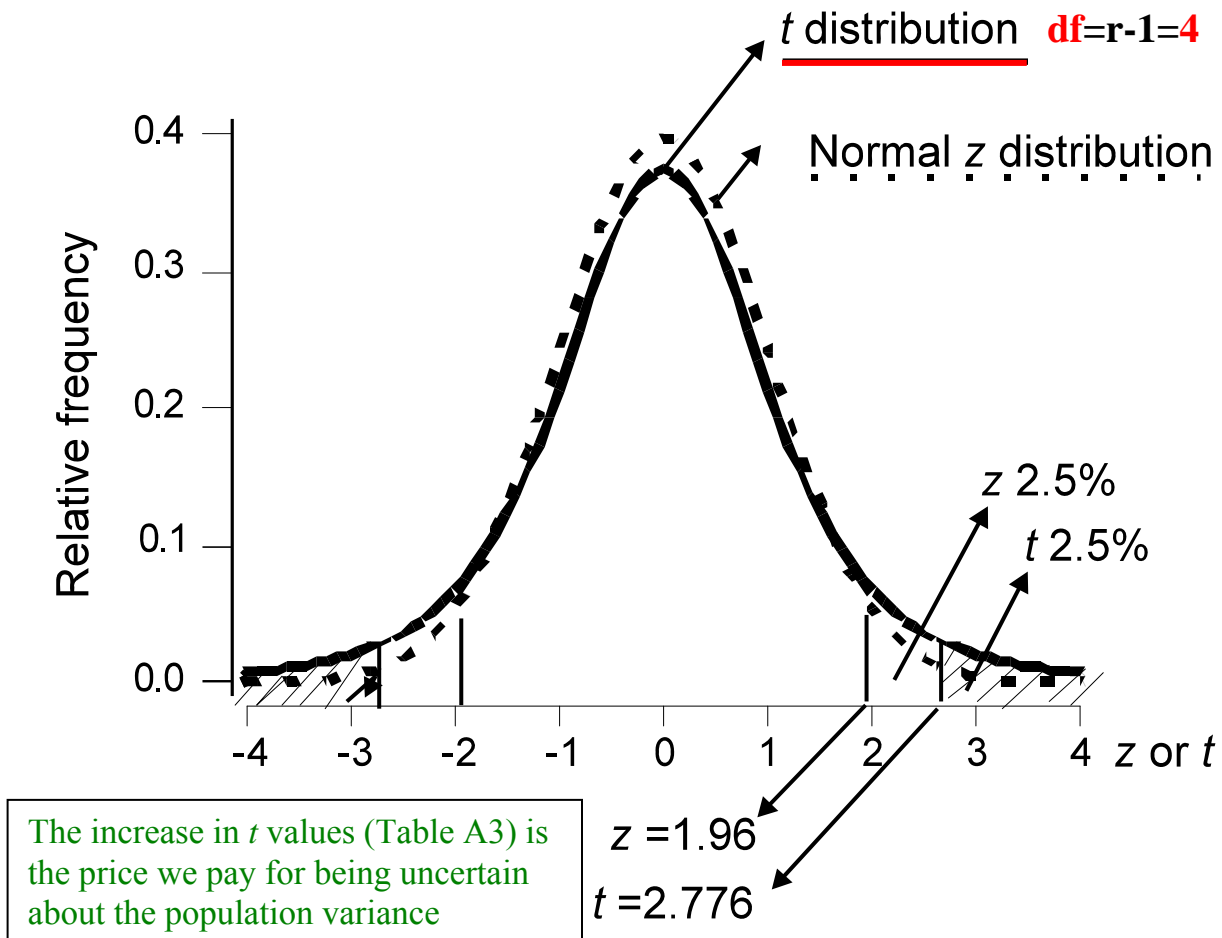


Fig. 1. Distribution of t ($df=4$) compared to z . The curve of t is symmetric and somewhat wider and flatter than the z distribution, lying under it at the center and above it in the tails.

2. 2. Confidence limits based on sample statistics

The general formula for any parameter δ is:

Estimated $\delta \pm \text{Critical value} * \text{Standard error of the estimated } \delta$

For \bar{Y} , that is $\bar{Y} \pm t_{\alpha/2, r-1} * s_{\bar{Y}}$

\bar{Y} is distributed about μ according to the t distribution so it satisfies

$$\Pr\{\bar{Y} - t_{\alpha/2, r-1} s_{\bar{Y}} \leq \mu \leq \bar{Y} + t_{\alpha/2, r-1} s_{\bar{Y}}\} = 1 - \alpha$$

For a confidence interval of size $1-\alpha$, use a t value corresponding to $\alpha/2$.

Therefore the confidence interval is

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

These two terms represent the lower and upper $1-\alpha$ **confidence limits** of the mean. The interval between these terms is called **confidence interval**.

Example: Data Set 1 of *Hordeum* 14 malt extraction values

$\bar{Y} = 75.94$ $s_{\bar{Y}} = 1.23 / \sqrt{14} = 0.3279$ A table gives the $t_{0.025, 13}$ value of 2.160,

95% confidence interval for $\mu = 75.94 \pm 2.160 * 0.3279 \Rightarrow [75.23- 76.65]$

If we repeatedly obtained samples of size 14 from the population and constructed these limits for each, we expect 95% of the intervals to contain the true mean.

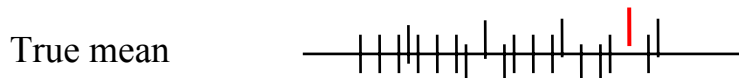


Fig. 2 Twenty 95% confidence intervals. One out of 20 intervals does not include the true mean.

2. 3. Hypothesis testing and power of the test.

Example Barley data. $\bar{Y} = 75.94$, $s_{\bar{Y}} = 0.3279$, $t_{0.025,13} = 2.160$, CI: [75.23- 76.65]

- 1) Test $H_0 \mu = 78$ against the $H_1 \mu \neq 78$.
- 2) Choose $\alpha < 0.05$
- 3) The formula for the test statistic t is as before:

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

For our example the sample value of the test statistic

$$t = (75.94 - 78) / 0.3279 = -6.3 \text{ (interpretation: 78 is 6.3 SE from the mean. Too far!).}$$

- 4) The absolute value of this statistics is > 2.16 ($t_{0.025,13}$) so we reject H_0 .

This is equivalent to calculate a 95% confidence interval for the mean. Since μ_0 (78) is not within the C.I. [75.23- 76.65] we reject H_0 .

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 (β)

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

Power of the test: The quantity $1 - \beta$ is the **power** of the test, and represents the probability of correctly rejecting a false null hypothesis. Is a measure of the ability of the test to detect μ_1 .

Note that for a given \bar{Y} and s , if 2 of the 3 quantities α , β , and r are specified then the third one can be determined. Choose the right number of replications to keep **Type I error $\alpha < 0.05$** and **Type II error $\beta < 0.20$** .

Power of a test (see ST&D pg 118-119 for origin of formula)

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

≠ between 2 means in SE units

What is the power of a test for $H_0: \mu = 74.88$ in the barley data set against $H_1: \mu = 75.94$. Since $\alpha = 0.05$, $r = 14$ ($t_{0.025,13} = 2.160$), and $s_{\bar{Y}} = 0.32795$.

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

Using the previous formula:

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

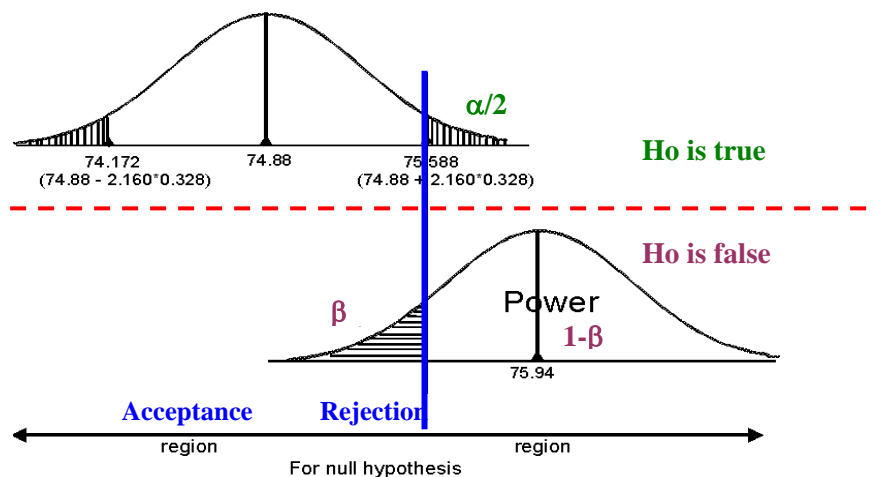


Fig. 3. Type I and Type II errors in the Barley data set.

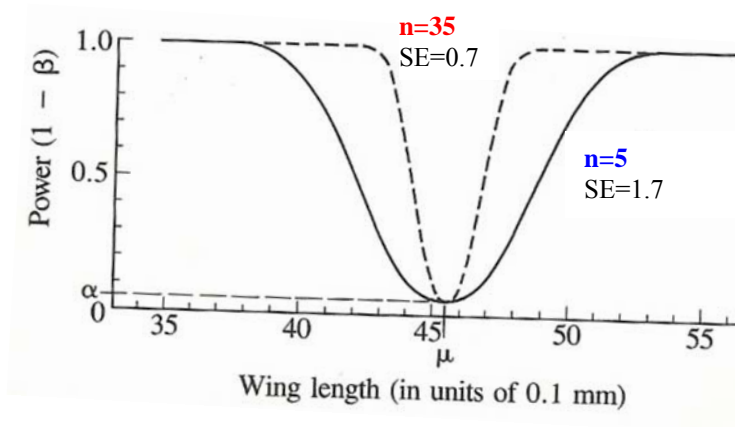
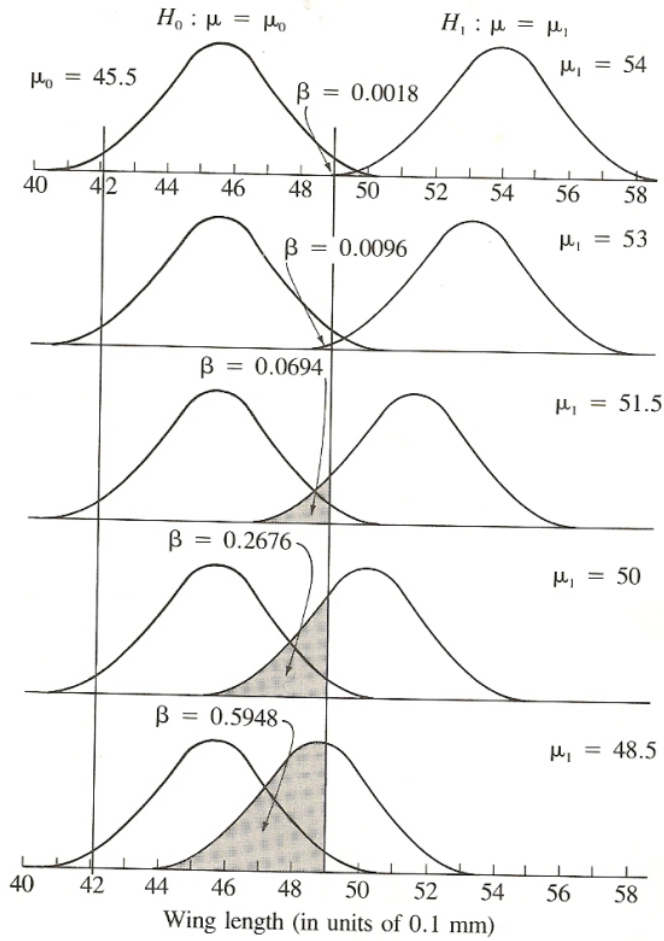
The area $1-\beta$ in the rejection region = the probability that $\mu > 75.588$ under H_1 = power =

$$P(t > (75.588 - 75.94) / 0.32795) = P(t > -1.072) = 0.85$$

The magnitude of β depends on **how far the alternative parametric mean** is from the parametric mean of the null hypothesis. As the alternative mean approaches the parametric mean, β increases to a maximum value of $1-\alpha$.

Power curves

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



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

Two types of alternative hypothesis

$H_0: \mu_1 - \mu_2 = 0$ versus $H_1: \mu_1 - \mu_2 \neq 0$ (two tail test) \rightarrow (t value: top t Table)
 $H_1: \mu_1 - \mu_2 > 0$ (one tail test) \rightarrow (t value: bottom t Table)

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

$$Power = P\left(t > t_{\frac{\alpha}{2}} - \frac{|\mu_1 - \mu_2|}{s_{\bar{Y}_1 - \bar{Y}_2}}\right) = P\left(t > t_{\frac{\alpha}{2}} - \frac{|\mu_1 - \mu_2|}{\sqrt{\frac{s_{pooled}^2}{N}}}\right),$$

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

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

When $n_1 = n_2 = n$ (equal sample sizes) that the formulas reduce to:

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

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

$$Power = P\left(t > t_{\frac{\alpha}{2}} - \frac{|\mu_1 - \mu_2|}{s_{\bar{Y}_1 - \bar{Y}_2}}\right) = P\left(t > t_{\frac{\alpha}{2}} - \frac{|\mu_1 - \mu_2|}{\sqrt{\frac{2s_{pooled}^2}{n}}}\right)$$

The variance of the difference between two random variables is the sum of the variances (error are always added) (ST&D 113-115).

The degrees of freedom for the critical $t_{\alpha/2}$ are

Equal sample size: $2*(n-1)$

Unequal sample size: $(n_1 - 1) + (n_2 - 1)$

2. 4. 2 Sample size for the estimation of the mean with known σ^2 . Using the z statistic

If the population variance σ^2 is known the z statistic for the standard normal distribution can be used. No initial sample is required to estimate the sample size. Recall that

$$Z = \frac{\bar{Y} - \mu}{\sigma} \quad \text{so} \quad \bar{Y} \pm Z_{\alpha/2} \sigma_{\bar{Y}} \quad \text{or} \quad \bar{Y} \pm Z_{\alpha/2} \frac{\sigma}{\sqrt{r}}$$

The formula for the confidence interval for the mean using the z statistic is

$$\left[\underbrace{\bar{Y}}_{d} \pm \underbrace{\sigma_{\bar{Y}}}_{d} \right] \quad d = z_{\alpha/2} \sigma / \sqrt{r} \quad d^2 = z_{\alpha/2}^2 \sigma^2 / r$$

With $d = \text{half-length}$ of the confidence interval. This can be rearranged to

$$r = z_{\alpha/2}^2 \sigma^2 / d^2 = z_{\alpha/2}^2 (\sigma/d)^2 = (1.96)^2 (\sigma/d)^2 = 3.8 (\sigma/d)^2$$

$$\text{So if } d = \sigma \Rightarrow r \approx 4 \quad d = 0.5 \sigma \Rightarrow r \approx 16 \quad d = 0.25 \sigma \Rightarrow r \approx 64$$

he equation may be also expressed in terms of the **coefficient of variation**, (use $CV = s / \bar{Y}$ as a proportion not as a %) and the population mean as

$$r = \frac{z_{\alpha/2}^2 CV^2}{(d/\mu)^2}$$

Here d/μ is the **confidence interval as a fraction of the population mean**. For example $d/\mu < 0.1$ means that the length of the confidence interval should not be larger than **two** tenth of the population mean.

$$d/\mu < 0.1 \quad \text{and so} \quad 2d/\mu < 0.2$$

Example:

The CVs of yield trials in our experimental station are never greater than 15%. How many replications are necessary to have a 95% CI for the true mean of less than 1/10 of the average yield?

$$2d = 0.1 \quad \text{so} \quad d = 0.1/2 = 0.05$$

$$r = 1.96^2 0.15^2 / 0.05^2 = 34.6 \approx 35$$

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

Unknown σ^2 . Stein's Two-Stage Sample

Consider a $(1 - \alpha)$ % confidence interval of a mean, μ .

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

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

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

This formula can be rearranged to estimate necessary sample size r

$$r = t_{\alpha/2, r-1}^2 s^2 / d^2 (\approx z^2 s^2 / d^2)$$

Stein's procedure is to use a pilot study to estimate s^2 and then compute r
We must iterate until the equation is satisfied with the same r values on both sides of the equals sign.

Example: 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 length of a 95% C.I. of the mean shorter than 5 cm?

Using $r = t_{\alpha/2, r-1}^2 s^2 / d^2$, r is estimated **iteratively**,

initial-n	$t_{5\%, df}$	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, he could have an estimated average height within 5 cm deviation of the true mean.

Note that if we start with “ z ” approximation, then

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

and then we can start the iterative process at 62, not too far from the exact estimate, 64.

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

In testing the hypothesis $H_0: \mu = \mu_0$, we can take into account the possibility of a **Type I** and **Type II** error **simultaneously**.

To calculate β we need to know the alternative μ_1 or at least the minimum difference we want to detect between the means $\delta = \mu_0 - \mu_1$.

The appropriate formula for computing r , the number of observations on **each** treatment, is given by equation:

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

For $\alpha = 0.05$ and $\beta = 0.20$, $z_{\beta} = 0.8416$ and $z_{\alpha/2} = 1.96$, $Z_{\alpha/2} + Z_{\beta} \approx 8$

If $\delta = 2\sigma \Rightarrow r \approx 4$
 If $\delta = 1\sigma \Rightarrow r \approx 16$
 If $\delta = \frac{1}{2}\sigma \Rightarrow r \approx 64$

The alternatives are two-tailed because we do not know if the alternative μ_1 is larger or smaller than μ_0 .

We rarely know σ^2 so we must estimate it with s :

$$r = 2 (s_p / \delta)^2 (t_{\alpha/2} + t_{\beta})^2 \quad \text{where } s_{\text{pooled}} = \text{SQRT}[(s_1^2 + s_2^2)/2]$$

the sample size, is estimated **iteratively**. 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:

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

We can also **define δ in terms of σ** . For example, we might want to detect a difference between means of one standard deviation in size.

Example: Two varieties will be compared for yield, with a previously estimated $s^2 = 2.25$ ($s=1.5$). How many replications should be used, so that these varieties can be compared with a $\alpha = 5\%$, and $\beta = 20\%$ to detect a difference greater than 1.5 tons/acre?

First: $r \approx 2 (\sigma/\delta)^2 (Z_{\alpha/2} + Z_{\beta})^2 = 2 (1.5/1.5)^2 (1.96+0.8416) = 15.7$

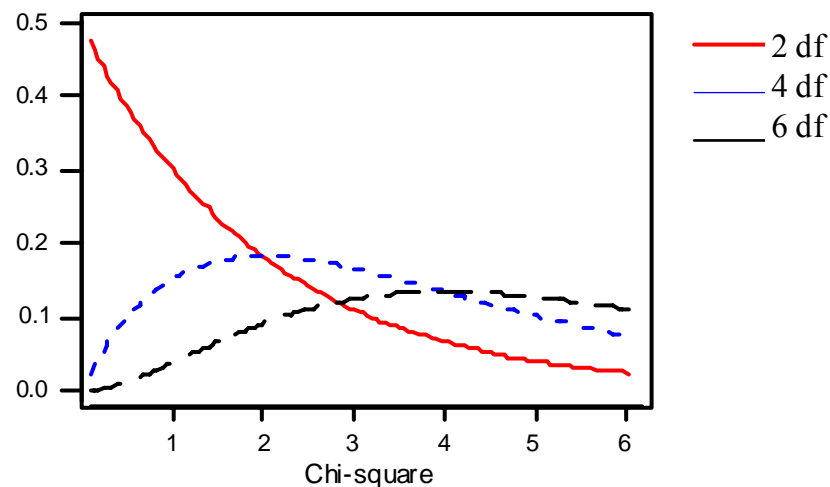
Then: $r = 2 (s / \delta)^2 (t_{\alpha/2} + t_{\beta})^2$ to estimate the sample size **iteratively**.

guesstimate n	df = 2(r - 1)	$t_{0.025}$	$t_{0.20}$	estimated n
16	30	2.0423	0.8538	16.8
17	32	2.0369	0.8530	16.7

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

The chi-squared distribution is used to establish confidence intervals around the sample variance as an estimate for the population variance.

2. 4. 5. 1. The Chi- square distribution



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

Relation between the normal and chi-square distributions.

If Z_1, Z_2, \dots, Z_n are random variables from a *standard* normal distribution then the sum $Z_1^2 + \dots + Z_n^2$ has a χ^2 distribution with n degrees of freedom.

$$\chi_{1}^2 = Z_{(0,1)}^2 = t_{\infty}^2 \quad \chi_{1, 0.05}^2 = 3.84 \text{ and } Z_{(0,1), 0.05}^2 = t_{\infty, 0.05}^2 = 1.96^2 = 3.84$$

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

When we change the parametric μ to a sample means this expression becomes

$$\frac{1}{\sigma^2} \sum (Y_i - \bar{Y})^2 = \frac{(n-1)s^2}{\sigma^2} \text{ that has a } \chi_{n-1}^2 \text{ distribution.}$$

2. 4. 5. 2. Confidence interval for σ^2

We can make the following statement about the ratio $(n-1) s^2/\sigma^2$ that has χ^2_{n-1} distribution,

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

Simple algebraic manipulation of the quantities in the inequality yields

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

or

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

The 1st form is particularly useful to construct 95% confidence interval for 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 .

Example: An experimenter wants to estimate σ with a 90% confidence that the estimated s is within 20% difference from σ , what is the required sample size?

$$\Pr \{ 0.8 \leq s/\sigma \leq 1.2 \} = 0.90 = \Pr \{ 0.64 \leq s^2/\sigma^2 \leq 1.44 \} = 0.90$$

thus

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

Since χ^2 is not symmetrical, the above two solutions may not be exactly equal if sample size is small. The actual computation involves the following iterative process:

guesstimate

size (n-1)	1 - $\alpha/2$ = 95%		$\alpha/2$ = 5%	
	$\chi^2_{(n-1)}$	$\chi^2_{(n-1)} / (n-1)$	$\chi^2_{(n-1)}$	$\chi^2_{(n-1)} / (n-1)$
20	10.90	0.545	31.4	1.57
30	18.50	0.616	43.8	1.46
40	26.50	0.662	55.8	1.40
35	22.46	0.642	49.8	1.42

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