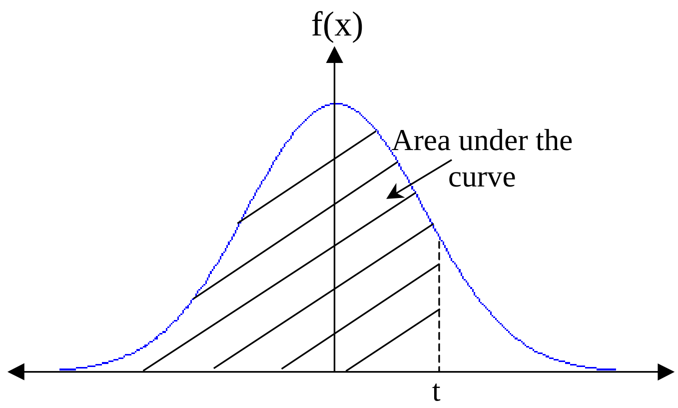
9. Statistical Analysis

Review: The one–dimensional Gaussian probability distribution function (PDF) of normally distributed random variable is:

$$f(x) = \frac{1}{\sigma_x \sqrt{2\pi}} \, e^{-\frac{(x-\mu_x)^2}{2\sigma_x^2}} ,$$

where μ and σ are the mean and standard deviation of variable (x). And the normal distribution function (NDF) as:



$$F(x) = \int_{-\infty}^{t} \frac{1}{\sigma_x \sqrt{2\pi}} \, e^{-\frac{(x-\mu_x)^2}{2\sigma_x^2}} \, dx$$

Where (t) is the upper bound of the integration.

- As stated before that the area under the curve represents the probability of occurrence. Furthermore, the integration of this function yields the area under the curve.
- Unfortunately, the integration of the above equation cannot be carried out in closed form, and thus numerical integration techniques must be used to tabulate values of this function. This has been done for the function when the mean is zero ($\mu = 0$) and the variance is 1 ($\sigma^2 = 1$).

The result of this integration is shown in the following table. The tabulated values represent the areas under the curve from –∞ to t.

t	Decimals of (t)									
	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
-3.2									.	
-3.1									.	
.									.	
.									.	
.									.	
-0.2									.	
-0.1									.	
0.0									.	
0.1									.	
0.2									.	
.									.	
1.6	.	.	.	.	.	.	.	.	**0.9535**	
.										
3.2										

Example:

- ✓ To determine the area under the curve from $-\infty$ to 1.68: Scan along row 1.6 and under column 0.8. At the intersection of row 1.6 and the column 0.8 (1.68), the value is 0.9535.

Meaning:

- Since the area under the curve represents probability, and its maximum area is 1, this means that there is 95.35% (0.95352= x 100) probability that (t) is less than or equal to 1.68.
- Alternatively, it can be stated that there is a 4.648% (1-0.9535) x 100 probability that (t) is greater than 1.68.

9.1. The Standard Normal Distribution Function

- ✓ Once available, this table can be used to evaluate the distribution function of any mean, μ, and variance σ^2.

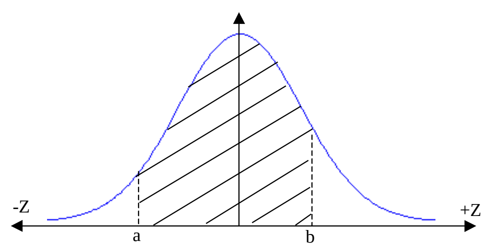
- ✓ For example, if y is a normal random variable with a mean, μ_y, and standard deviation σ_y, an equivalent normal random variable $z = (y - \mu_y)/\sigma_y$ can be defined that has $\mu = 0$ and $\sigma^2 = 1$.

- ✓ Substituting the definition for Z with $\mu_z = 0$ and $\sigma_z^2 = 1$ into the NDF

$$N(Z) = \int_{-\infty}^{t} \frac{1}{\sqrt{2\pi}} e^{-\frac{Z^2}{2}} dZ$$

This is known as the Standard Normal Distribution Function.

Characteristics of the NDF



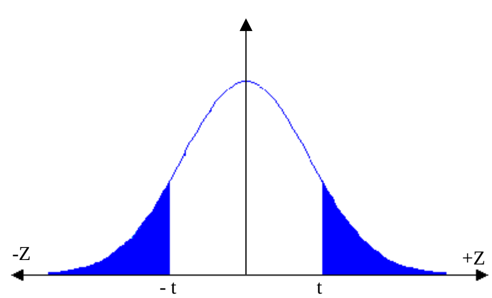
- $P(Z < t)$ = **N(t)**
- $P(a < Z < b)$ = **N(b) – N(a)**

- If $-a = b$ = t
  - $P(-t < Z < t) = P(|Z| < t)$ = **N(t) – N(-t)**

- From symmetry of the normal distribution, it is seen that:

  - $P(Z > t)$ = $P(Z < -t)$
  - $1 - N(t)$ = $N(-t)$

Therefore, $P(|Z| < t) = N(t) - N(-t)$ = **2 N(t) – 1**



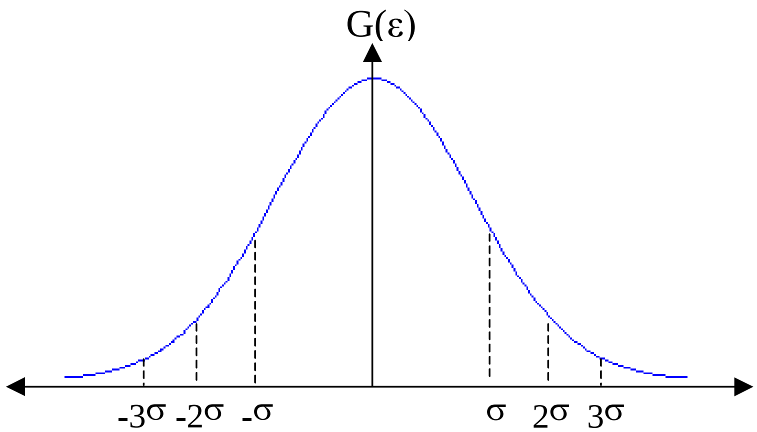
Probability of the Standard Deviation (Standard Error)

From the above characteristic, the probability of the standard deviation can be estimated as follows:

$$P(-\sigma < Z < \sigma) = N(\sigma) - N(-\sigma) \quad \ldots\ldots\ldots \quad \text{Recall that } \sigma = 1$$

Looking into the table for t = 1 and t = -1, the area between $-\sigma$ and σ is:

$$P(-\sigma < Z < \sigma) = 0.84134 - 0.15866 = 0.68268 \ \text{(which also equal} = 2 \times 0.84134 - 1)$$



$$\left.\begin{array}{r} P(-\sigma \le \varepsilon \le \sigma) = 0.683 \\ P(-2\sigma \le \varepsilon \le 2\sigma) = 0.954 \\ P(-3\sigma \le \varepsilon \le 3\sigma) = 0.997 \end{array}\right\} \textbf{called confidence intervals}$$

<u>Meaning:</u>

For any group of measurements there is approximately a 68.3 % chance that any single observation has an error $\pm\sigma$. This is true for any set of measurements having normally distributed errors.

9.2. The 50% Probable Error:

For any group of observations, the 50% probable error establishes the limits within which 50% of the errors should fall.

In other words, any measurement has the same chance of becoming within these limits as it has as falling outside them.

$$P(|Z| < t) = N(t) - N(-t) = 0.5$$
$$= 2N(t) - 1 = 0.5$$

Therefore : $1.5 = 2N(t)$ or $0.75 = N(t)$

From the SNDF tables it is apparent that 0.75 is between t value of 0.67 and 0.68, that is
$N(0.67) = 0.7487$ and $N(0.68) = 0.7517$

The t value can be found by linear interpolation, as follows:

$$\frac{\Delta t}{0.68-0.67} = \frac{0.7500-0.7486}{0.7517-0.7486} = \frac{0.0014}{0.0031} = 0.4516$$

or $\Delta t = 0.01 \times 0.4516$

and therefore, $t = 0.67 + 0.0045 = 0.6745$

For any set of observations, therefore, the 50% probable error can be obtained by computing the standard error and then multiplying it by 0.6745, i.e. $\boxed{E_{50} = \pm 0.6745\,\sigma}$

9.3. The 95% Probable Error:

- The 95% probable error, or E95, is the bound within which, theoretically, 95% of the observation group's errors should fall.

- The error category is popular with geomatics engineers for expressing precision.
- Using the same reasoning as in the in the equations for the 50% probable error:

  $$P(|Z| < t) = N(t) - N(-t) = 0.95$$
  $$= 2N(t) - 1 = 0.95$$

  Therefore :
  $1.95 = 2N(t)$ or $0.975 = N(t)$

- From the SNDF tables it is determined that 0.975 occurs with a (t) value of 1.96.
- Thus, for any set of observations, the 95% probable error can be obtained by computing the standard deviation (or error) and then multiplying it by 1.96, i.e.

$$E_{95} = \pm 1.96\,\sigma$$

9.4. Other Percent Probable Errors

Using the same computation procedures, other percent error probable can be calculated.

Probable Error	Multiplier
E_{50}	0.6745σ
E_{90}	$1.6449\ \sigma$
E_{95}	$1.96\ \sigma$
E_{99}	$2.576\ \sigma$
$E_{99.7}$	$2.965\ \sigma$
$E_{99.9}$	$3.29\ \sigma$

The $E_{99.7}$ (or a similar percent probable error) is often used in detecting blunders as will be explained in the following example.

Example:

The arc-second portion of 50 direction readings from 1″ instrument are listed below. Find the mean, standard deviation (error), and the E_{95}. Check the observations at a 99% level of certainty for blunders.

41.9	46.3	44.6	46.1	42.5	45.9	45.0	42.0	47.5	43.2	43.0	45.7	47.6
49.5	45.5	43.3	42.6	44.3	46.1	45.6	52.0	45.5	43.4	42.2	44.3	44.1
42.6	47.2	47.4	44.7	44.2	46.3	49.5	46.0	44.3	42.8	47.1	44.7	45.6
45.5	43.4	45.5	43.1	46.1	43.6	41.8	44.7	46.2	43.2	46.8		

Solution:

The mean = $\frac{\sum_{i=1}^{50} l_i}{n} = \frac{2252}{50} = 45.04''$ & the standard deviation = $\sqrt{\frac{\sum_{i=1}^{50}(Mean - v_i)^2}{50-1}} = \pm 2.12''$

$\therefore$ $E_{95} = \pm 1.96\ \sigma = \pm 4.16''$ (Thus 95% of the data should fall between 45.04 ± 4.16″ or in the “40.88 - 49.20” range)

The data actually contain three values that deviate from the mean by more than 4.16 (i.e. that are outside the range 40.88 to 49.20). They are 49.5 (2 times) and 52.0. No values are less than 40.88, and therefore 47/50 = 94% or the measurements lie in the E_{95} range.

$\therefore$ $E_{99} = \pm 2.576\ \sigma = \pm 5.46''$ (Thus 99% of the data should fall between 45.04 ± 5.46″ or in the “39.58 - 50.50” range).

Actually, one value is greater than 50.50, and thus 49/50 = 98% of all measurements fall in this range.

9.5 Confidence Intervals and Statistical Testing

- The following contains a finite population of 100 values. The mean (μ) and the variance (σ^2) of that population are 26.1 and 17.5, respectively.

18.2	26.4	20.1	29.9	29.8	26.6	26.2
25.7	25.2	26.3	26.7	30.6	22.6	22.3
30.0	26.5	28.1	25.6	20.3	35.5	22.9
30.7	32.2	22.2	29.2	26.1	26.8	25.3
24.3	24.4	29.0	25.0	29.9	25.2	20.8
29.0	21.9	25.4	27.3	23.4	38.2	22.6
28.0	24.0	19.4	27.0	32.0	27.3	15.3
26.5	31.5	28.0	22.4	23.4	21.2	27.7
27.1	27.0	25.2	24.0	24.5	23.8	28.2
26.8	27.7	39.8	19.8	29.3	28.5	24.7
22.0	18.4	26.4	24.2	29.9	21.8	36.0
21.3	28.8	22.8	28.5	30.9	19.1	28.1
30.3	26.5	26.9	26.6	28.2	24.2	25.5
30.2	18.9	28.9	27.6	19.6	27.9	24.9
21.3	26.7					

- By **randomly selecting 10 values** of this table, an estimate of the mean and the variance ($\bar{X}$ and S^2) of this **sample** can be estimated.
- However, it would not be expected that these estimates ($\bar{X}$ and S^2) exactly match the mean and the variance of the population. Now if the sample size were increased, it would be expected that the $\bar{X}$ and S^2 would better match μ and σ^2 as shown in the table below.

Increasing Sample Sizes

No.	$\bar{X}$	S^2
10	26.9	28.1
20	25.9	21.9
30	25.9	20.0
40	26.5	18.6
50	26.6	20.0
60	26.4	17.6
70	26.3	17.1
80	26.3	18.4
90	26.3	17.8
100	26.1	17.5

- Since the mean and the variance of the sample ($\bar{X}$ and S^2) are computed from random variable, they are also random variables. Thus it is concluded that the values computed contain errors.

Random sample sets from population ($\mu = 26.1$ and $\sigma^2 = 17.5$).

Set 1:	29.9, 18.2, 30.7, 24.4, 36.0, 25.6, 26.5, 29.9, 19.6, 27.9	$\bar{X} = 26.9$, $S^2 = 28.1$
Set 2:	26.9, 28.1, 29.2, 26.2, 30.0, 27.1, 26.5, 30.6, 28.5, 25.5	$\bar{X} = 27.9$, $S^2 = 2.9$
Set 3:	32.2, 22.2, 23.4, 27.9, 27.0, 28.9, 22.6, 27.7, 30.6, 26.9	$\bar{X} = 26.9$, $S^2 = 10.9$
Set 4:	24.2, 36.0, 18.2, 24.3, 24.0, 28.9, 28.8, 30.2, 28.1, 29.0	$\bar{X} = 27.2$, $S^2 = 23.0$

- **Fluctuations in the mean and variance** computed from sample sets **raises questions about the reliability of these estimates**.
- A higher confidence value is likely to be placed on a sample set with small variance than on one with large variance.
- In the above table, because of its small variance, one is more likely to believe that the mean of the 2nd set is the most reliable estimate of the population mean. In reality this is not the case, as the means of the other three sets are actually closer to the population mean ($\mu = 26.1$).
- The estimation of the mean and variance of a variable from sample data is referred as **point estimation** (because it results in one value for each parameter in question)
- After having performed a point estimation, **the question remains as how much the deviation** of the estimate is likely to be from the still unknown true values of the parameters (mean and variance). In other word, we would like to have an indication of how good the estimation is and how much it can be relied on.
- **An absolute answer to this question is not possible because sampling never leads to the true parameters**.
- It is only possible to estimate probabilities with which the true value of the parameters in question is likely to be within a certain interval around the estimate. Such probabilities can be determined if the distribution function F(X) of the random variable is given.

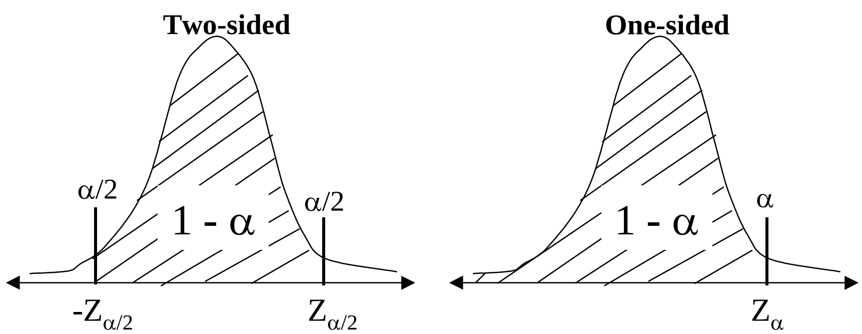
Recall: The probability that a random variable Z takes values within the boundary "a and b" is given by:

$$P(a < Z < b) \quad = \quad F(b) - F(a) = \int_{x=a}^{x=b} f(x)\ dx$$

- By analogy to this, the probability statement for a **confidence interval** of a parameter Z, the estimate of which is $\hat{Z}$, is: $\mathbf{P(a < Z < b) \;=\; 1 - \alpha}$
- Where $(1 - \alpha)$ is called the ***confidence level*** or degree of confidence which is conventionally taken to be 90%, 95%, or 99%. And the values of **"a and b"** are called the **"*upper and lower confidence limits*"** for the parameter Z.
- The probability that the parameter does not fall in a given interval is α.

9.5. Confidence Intervals for Means

- **Given:** $\bar{X}$, σ^2, and n (number of observation) ……. (i.e. σ is known)
- **Required**: the (**1 – α**) confidence interval for μ (which is ~~un~~known).
- Recall: $\dfrac{\bar{X}-\mu}{\sigma/\sqrt{n}}$ is normally distributed with mean 0.0 and variance 1.0
- The probability statement for the confidence interval, which is symmetric here, is then:



$$P\left\{-Z_{\alpha/2}<\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}<Z_{\alpha/2}\right\}=1-\alpha$$

or

$$P\left\{\bar{X}-Z_{\alpha/2}\cdot\sigma/\sqrt{n}<\quad\mu\quad<\bar{X}+Z_{\alpha/2}\cdot\sigma/\sqrt{n}\right\}=1-\alpha$$

- For example for α = 5%, $Z_{1-\alpha/2}$ = 1.96, therefore we can write the above equation as:

$$P\left\{\bar{X}-1.96\sigma/\sqrt{n}<\quad\mu\quad<\bar{X}+1.96\sigma/\sqrt{n}\right\}=0.95$$

- The above is an example of the so called two-sided confidence interval. On the other hand, for a one-sided confidence interval we write:

$$P\left\{\mu\quad<\bar{X}+Z_{\alpha}(\sigma/\sqrt{n})\right\}=1-\alpha$$

- Therefore for α = 5%,

$$P\left\{\mu\quad<\bar{X}+1.6449\sigma/\sqrt{n}\right\}=0.95$$

Example:

Suppose a distance is measured **n = 8** times with the mean $\bar{X}$ **= 10.1 cm**. We assume that the variance of the normal population is known to be $\boldsymbol{\sigma^2 = 0.1}$ **cm²**. Then the **95% confidence interval on μ** (which is unknown) for the two-sided confidence interval is:

$$P\left\{\bar{X} - Z_{\alpha/2} \cdot \sigma/\sqrt{n} < \mu < \bar{X} + Z_{\alpha/2} \cdot \sigma/\sqrt{n}\right\} = 1 - \alpha$$

$$P\left\{10.1 - 1.96\sqrt{\frac{0.1}{8}} < \mu < 10.1 + 1.96\sqrt{\frac{0.1}{8}}\right\} = 0.95$$

$$P\{9.88 < \mu < 10.32\} = 0.95$$

For a one sided interval,

$$P\left\{\mu < \bar{X} + 1.6449\,\sigma/\sqrt{n}\right\} = 0.95$$

$$P\left\{\mu < 10.1 + 1.6449\sqrt{\frac{0.1}{8}}\right\} = P\{\mu < 10.28\} = 0.95$$

- Let us now consider the case in which the standard deviation of the distribution σ is not known and has to be replaced by the standard deviation of the sample, S. Therefore, the estimator under question is $\frac{\bar{X} - \mu}{S/\sqrt{n}}$ which has a different distribution than $\frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$.

9.6. Some Often Used Distributions

In connection with the theory of errors of observations and least squares adjustment, there are a few (one-dimensional) distributions that are often used. Only continuous distributions are discussed, particularly those used for statistical testing.

9.6.1. Gaussian or Normal Distributions

The Gaussian or ND is the most frequently used distribution in statistical theory. Its density function is given by:

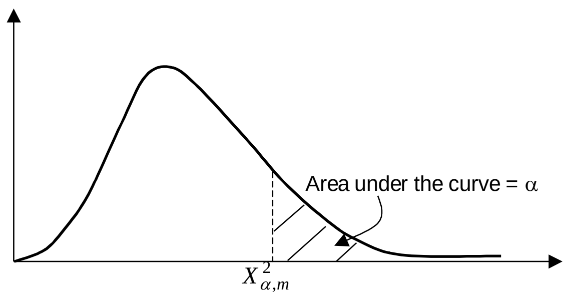
$$f(x) = \frac{1}{\sigma_x \sqrt{2\pi}}\, e^{-\frac{(x - \mu_x)^2}{2\sigma_x^2}}$$

which is fully described by to parameters μ_x and σ_x.

The cumulative NDF of the standardized random variable $Z = \frac{X - \mu_X}{\sigma_X}$ (having a zero mean and unit standard deviation, i.e. $[\mu_z = 0 \,\&\, \sigma_z = 1]$) is given by:

$$F(Z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{Z} e^{\frac{-Z^2}{2}} dz$$; note that the F(Z) value can be extracted from a table

9.6.2. The χ^2 distribution (chi square)



- The distribution of the sum of squares of independent random variables each of which is normally distributed is known as χ^2 (chi square) distribution.

Examples:

- $\chi_n^2 = X_1^2 + X_2^2 + \ldots\ldots\ldots\ldots + X_n^2$

  which is χ_n^2 distribution of "n degrees of freedom"
- $v_{n,1}^T P_{n,n} v_{n,1}$ is a χ_n^2 distribution of "n degrees of freedom"

Entry to χ^2-distribution tables:

$\chi^2_{\alpha,m}$: α = significance region = area under the curve (to the right of the Chi-squared value);
1-α = confidence region = area under the curve (to the left of the Chi-squared vale)
m = degrees of freedom

m \ α	0.995	0.99	.. ..	0.0005
1				
2				
120				

$$P\left(\chi_m^2 > \chi_{\alpha,m}^2\right) = \alpha$$

$$= \int_{\chi^2_{\alpha,m}}^{\infty} f\left(\chi^2\right) d\chi^2$$

$$= 1 - \int_{0}^{\chi^2_{\alpha,m}} f\left(\chi^2\right) d\chi^2$$

9.6.3. The F (Fisher) Distribution

- The distribution of the ratio of two independent random variables, each having a χ^2 (chi-square) distribution, is an F-distribution.

  $$F_{m,n} = \frac{\chi_m^2 / m}{\chi_n^2 / n}$$ is F-distribution of (m) and (n) degrees of freedom.

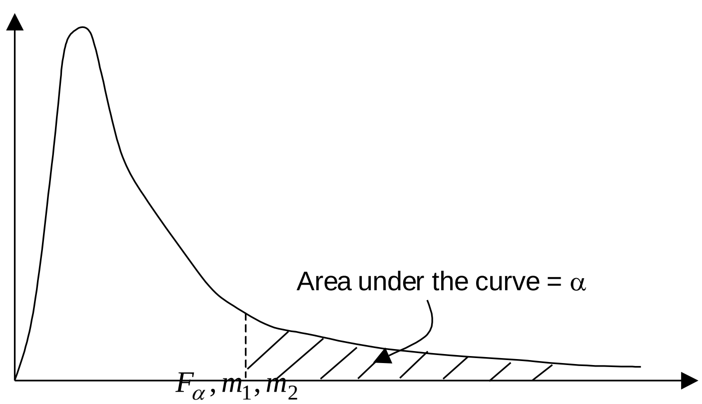
- The practical application of the F distribution in least squares adjustment and statistical testing is concerned with the comparisons of variances such as those obtained from two adjustments.
- In some practical cases the comparison may be between a variance obtained from the same adjustment (such as $\hat{\sigma}_0^2$) and an *a priori* given reference variance (such as σ_0^2), this case refers to the $F_{r,\infty} = \chi_r^2$ (where r is the degrees of freedom).

Recall: variance $= \sigma^2 = \dfrac{\sum v^2}{n-1}$, where $\sum v^2$ *is* χ_{n-1}^2 distribution

F_α, m_1, m_2

A number of tables have been developed for different α

e.g. F0.05, m_1, m_2



$m_1 \backslash m_2$	1	2	...	16
1				
2				
3				
.				
.				
.				
100				
.				
∞				

$$P\left(F_{m_1 m_2} > F_{0.05,m_1,m_2}\right) = 0.05$$

$$= \int_{F_{0.05,m1,m2}}^{\infty} f(F)\,dF$$

$$= 1 - \int_0^{F_{0.05,m1,m2}} f(F)\,dF$$

9.6.4. The t (student) Distribution

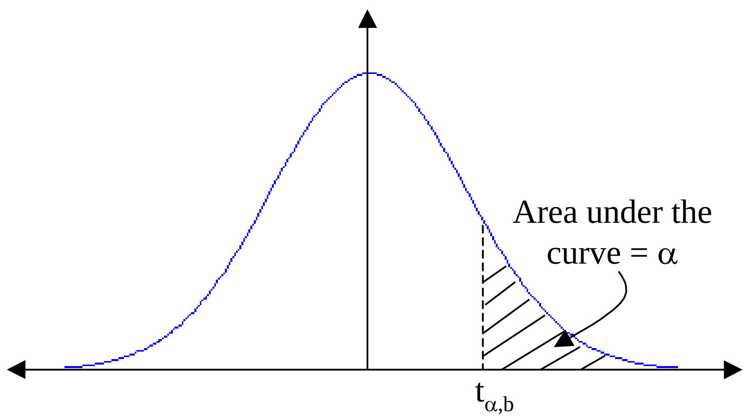
The t-distribution is used in connection with sampling (i.e. testing using sample statistics instead of population parameters)

Let $X_1', X_2', \ldots\ldots X_n'$ be n independent random (stochastic) variables of identical normal distribution with mean μ and standard deviation σ. Then the random variable:

$t = \dfrac{\bar{x} - \mu}{s}\sqrt{n}$ is said to have a t-distribution with (n-1) degrees of freedom, where

$\bar{x} = \dfrac{1}{n}\sum_{i=1}^{n} x_i \; and \; s^2 = \dfrac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$, where $\bar{x}$ **and** s^2 are the sample mean and standard deviation, respectively.

t-distribution table ($t_{\alpha,m}$)



m\α	0.25	0.2	…	0.0005
1				
2				
.				
.				
.				
120				

$$P(t_m > t_{\alpha,m}) = \alpha$$

$$= \int_{t_{\alpha,m}}^{\alpha} f(t)dt$$

$$= 1 - \int_{-\infty}^{t_{\alpha,m}} f(t)\; dt$$

Examples on Confidence Intervals

1. <u>Confidence Intervals for a Population Variance:</u>

- The χ^2 distribution is used in sampling statistics to determine the range within which the variance of the population can be excepted to occur based on:
  1) Some specified percentage probability
  2) The variance of a sample set
  3) The number of degrees of freedom of the sample

- Suppose you are given a random sample of variance s^2 from a normal population with variance σ^2.

- **<u>The Random Variable</u>:** $\frac{ms^2}{\sigma^2}$ has a χ^2 distribution with (m) degrees of freedom (where m is degrees of freedom used in computing s^2.)

- For (1-α) confidence level, this variable will take a value between

- $P\{\chi^2 > \chi^2_{\alpha,m}\} = \alpha$ **where** $\chi^2 = \frac{mS^2}{\sigma^2}$

- Unlike the normal distribution curve (and also the t-distribution), the χ^2 distribution (and also the F-distribution) is not symmetric about zero. To locate an area under the lower tail of the curve, the appropriate value of $\chi^2_{1-\alpha}$ must be found, where $P\{\chi^2 > \chi^2_{1-\alpha,m}\} = 1-\alpha$. These facts are used to construct the probability statement for χ^2 (which equal $\frac{ms^2}{\sigma^2}$ in this case) as follows:

$$P\left\{\chi^2_{(1-\frac{\alpha}{2}),m} < \frac{ms^2}{\sigma^2} < \chi^2_{\frac{\alpha}{2},m}\right\} = 1-\alpha$$

or

$$P\left\{\frac{ms^2}{\chi^2_{\frac{\alpha}{2}}} < \sigma^2 < \frac{ms^2}{\chi^2_{1-\frac{\alpha}{2}}}\right\} = 1-\alpha$$

For instance with m=10 and α = 0.05, we get (from the χ^2 table)

$$\chi^2_{0.025,10} = 20.48 \text{ and } \chi^2_{0.975,10} = 3.25$$

Note: This interval is two sided but is <u>asymmetric (not symmetric)</u>

For the one sided interval we have $\chi^2_{0.05,10} = 18.31$

Example: Suppose, as in the previous example n=8, $\overline{X}$ = 10.1 cm, and S^2 = 0.10 cm^2. The 99% confidence interval on the population variance σ^2 is: (note: m = n-1 = 8-1 = 7 and α = 0.01)

$$P\left\{\frac{(7)(0.1)}{\chi^2_{0.005,7}} < \sigma^2 < \frac{(7)(0.10)}{\chi^2_{0.995,7}}\right\} = 0.99$$

$$\therefore P\left\{\frac{(7)(0.1)}{20.28} < \sigma^2 < \frac{(7)(0.1)}{0.99}\right\} = 0.99$$

$$or\ P\left\{0.035cm^2 < \sigma^2 < 0.707cm^2\right\} = 0.99$$

Which state that the probability is 0.99 that the population variance lies between 0.035-0.707 cm^2

- Note:

$$P\{0.186cm < \sigma < 0.841cm\} = 0.99$$

σ = standard deviation

2. Confidence Intervals for Ratios of Two Variances:

- Another common statistical testing example is the comparison of two population variances. If we are given two independent random samples of sizes m_1 and m_2 from normal populations with variances σ_1^2 and σ_2^2, respectively, then each of the random variables $\frac{m_1 S_1^2}{\sigma_1^2}$ and $\frac{m_2 S_2^2}{\sigma_2^2}$ has a χ^2 distributions with m_1 and m_2 degrees of freedom, respectively. The ratio:

$$F_{m_1,m_2} = \frac{\chi^2_{m_1}/m_1}{\chi^2_{m_2}/m_2}$$

- Substituting for the $\chi^2_{m_1}$ **and** $\chi^2_{m_2}$ and reducing gives:

  $F = \dfrac{(m_1 S_1^2/\sigma_1^2)/m_1}{(m_2 S_2^2/\sigma_2^2)/m_2} = \dfrac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2}$ which is an F-distribution with m_1 and m_2 degrees of freedom.

- To establish a confidence interval for the ratio, the lower and upper values corresponding to the distribution's tails must be found. Similarly, to the χ^2 distribution, the F-distribution is not symmetric about zero. To locate an area under the lower tail of the curve, the appropriate value of $F_{1-\alpha}$ must be found through the following relation (this means we use the same F table):

$$\boxed{F_{(1-\alpha),m1,m2} = \frac{1}{F_{\alpha,m2,m1}}}$$

- A probability statement for the confidence interval for the ratio is constructed as follows:

$$P\left\{F_{(1-\alpha/2),m1,m2} < \frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2} < F_{(\alpha/2,m1,m2)}\right\} = 1-\alpha$$

or

$$P\left\{\frac{S_2^2}{S_1^2}F_{1-\alpha/2,m1m2} < \frac{\sigma_2^2}{\sigma_1^2} < \frac{S_2^2}{S_1^2}F_{\alpha/2,m1,m2}\right\} = 1-\alpha$$

$$Note: \quad F_{1-\alpha/2,m1,m2} = \frac{1}{F_{\alpha/2,m2,m1}}$$

For $m_1 = 10$, $m_2 = 60$, and $\alpha = 0.05$ (use the curve for $\alpha/2$)

$$F_{0.975,10,60} = \frac{1}{F_{0.025,60,10}} = \frac{1}{3.20} = 0.312$$

$$F_{0.025,10,60} = 2.27$$

Which provides the following confidence interval for $\frac{\sigma_2^2}{\sigma_1^2}$

$$0.312 \; \frac{S_2^2}{S_1^2} < \frac{\sigma_2^2}{\sigma_1^2} < 2.27\frac{S_2^2}{S_1^2}$$

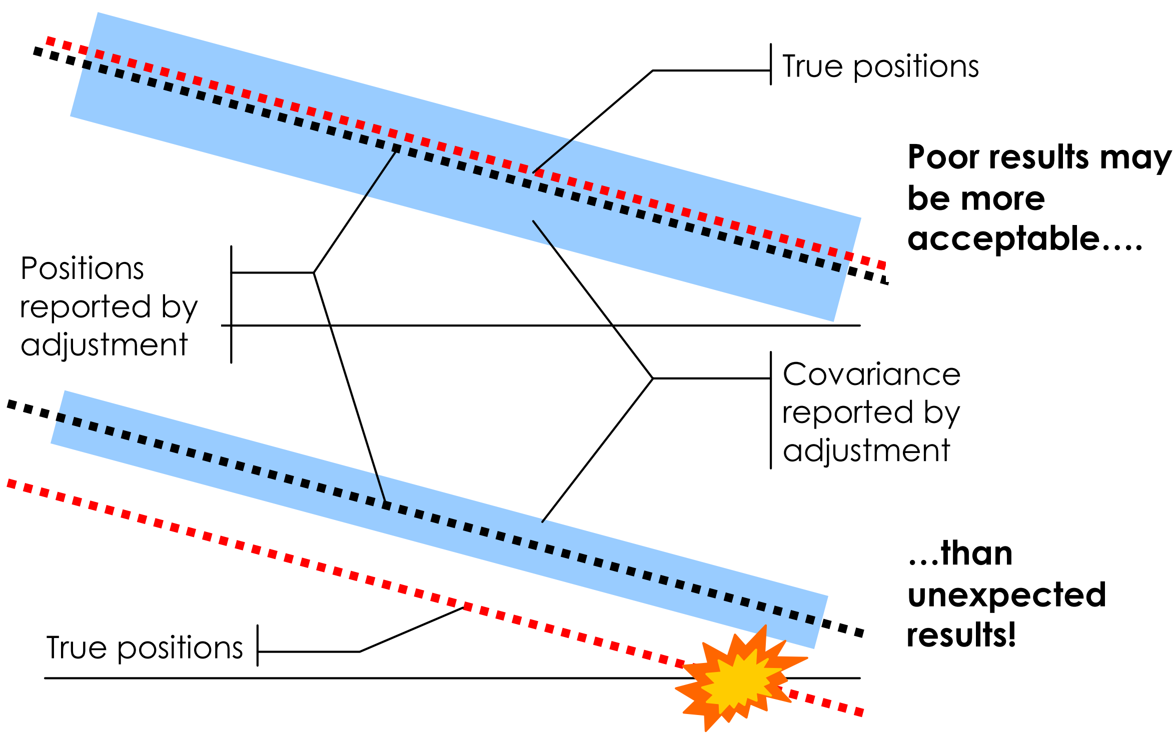
9.7. Statistical Testing

Fundamentals of Statistical Testing

- **Purpose:** To test the validity of statistics estimated using least squares (LS)
- **Basic Postulate:** The process of observation has been modeled as random variable
- **Note:** Although LS does not require that the PDF of the observations be known, the PDF is required for post-adjustment testing.

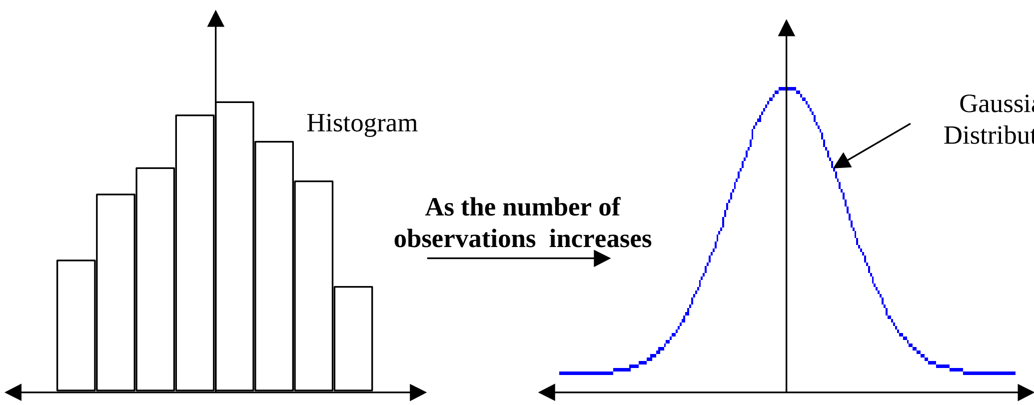
Motivation

- In many applications, the reliability of the adjustment results is at least as important as the actual results themselves.
- In other words, it is vital that the true errors in the adjusted parameters are consistent with the covariance information output by the adjustment.
- For example, in guiding an aircraft on its final approach to a runway, it is critical that a pilot be able to trust the positions reported by the GPS navigation system. If the positions are accurate at about the metre level, then this is acceptable. However, if positions are reported as being accurate to the nearest centimetre but are in fact several metres in error, disaster may occur.



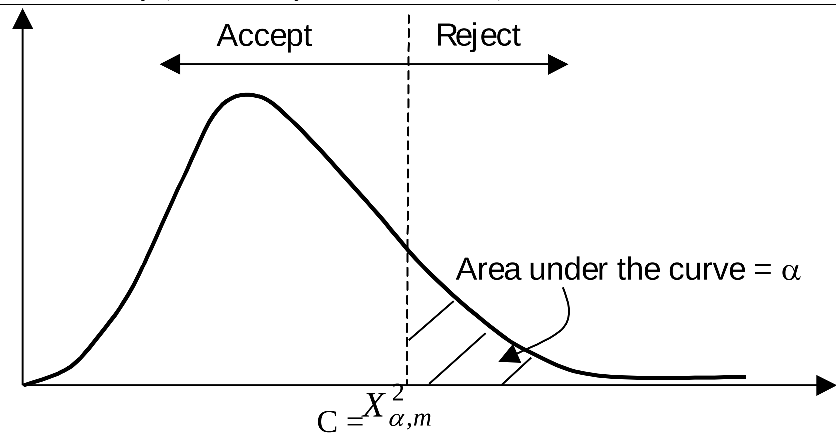
<u>Fundamentals of Statistical Testing</u>

- In hypothesis (guess or suggestion) testing, an assumption is implicitly or explicitly made about the PDF of a random variable (RV). The validity of that assumption is then tested.
- Let us consider the case of a sample of (n) observations of the length of a baseline. If we draw the "histogram" of these observations we see, as we obtain more and more observations, that the histogram tends to take the general shape of the "Gaussian distribution".

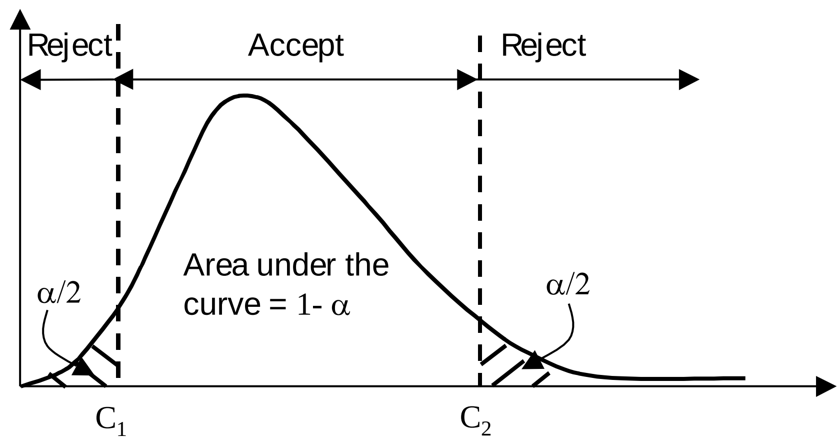


- On these bases, we may postulate (guess/suggest) that our observations have a Gaussian distribution with a certain mean (or average) and certain variance. We call the mean and variance "parameters" of the distribution (recall: the Gaussian distribution is completely specified by these two parameters).
- The mathematical abstractions we are making here can be interpreted as follows: we are representing the process of measuring the length of the baseline by a "random variable", for example ℓ whose distribution is $\phi(\ell)$. Where a random variable is simply a variable whose probability of taking on a specific value is given by its PDF.
- Such suggestion is known as <u>"the basic postulate of mathematical statistics"</u> and is the basis for all our statistical testing procedures.
- Let us now make different assumptions or "hypothesis" about $\phi(\ell)$ and its parameters. For example, **<u>let us assume (or hypothesize) that $\phi(\ell)$ is a normal distribution</u>** (perhaps based on previous experience). This basic assumption, which we wish to test is called the "null hypothesis" is denoted by Ho.

We now want to devise a procedure to test (statistically) the null hypothesis on the basis of our collected measurements. That is, **<u>we seek to confirm or reject the assumption that our measurements are normally distributed</u>**. The testing procedure we use in this case is the chi-square (x^2) goodness of fit test (because we are testing how $\phi(\ell)$ is close to the normal distribution). Using this test, we compute a "statistic" y on the basis of our measurements and compare it with a value C which we obtain from tables of the x^2 distribution. **<u>If y < C, we have (statistical) evidence that our null hypothesis should be accepted' if y > C, we have evidence that H_o should be rejected</u>**.



In the above example, we applied a "one-sided" statistical test. It is more common to use the "two-sided" statistical test which results in computing "critical values" C_1 and C_2 between which our statistic should lie if our null hypothesis is to be accepted.



Steps in hypothesis (assumption) testing:

1. Formulate a null hypothesis (H_0) about the PDF of an RV. Also formulate an alternative hypothesis (H_a) to accept if H_0 fails

e.g. $\mathbf{H_0: \mu =} \ \bar{X}$ $\mathbf{H_a: \mu \neq} \ \bar{X}$,

which states that the population mean is the same the sample mean

2. Formulate or compute a test statistic (e.g. for the above example the test statistics is y = $\frac{\bar{X}-\mu}{S/\sqrt{n}}$)

3. Decide on the PDF of the test statistic (i.e. decide on which distribution "t-distribution, χ^2 or F distribution" corresponds to the test statistic; e.g. for the above example the test statistic has a t-distribution). This will depend on the nature of the quantity being tested, for example:

- Are the population or sample parameters available?
- Is the statistics a summation of the squares of normally distributed random variables, then use χ^2 distribution as the PDF.
- Is the statistics is a ratio of two χ^2 distributed random variables, then use F-distribution as the PDF.

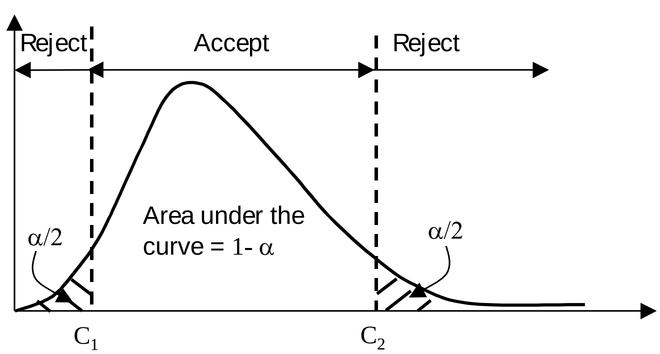
4. Make a probability statement that if the probability that the test statistic (an RV) lies between two critical values it is equal to the confidence level of the test.

$$P\left[C_1 < y < C_2\right] = 1 - \alpha \qquad \text{y = test statistics}$$

C_i are the critical values from the PDF tables

5. Make a decision from the confidence interval

- If the tested quantity (y) lies within the confidence interval (i.e. between C_1 and C_1) then accept H_0



- If tested quantity (y) lies outside the interval (C_1 and C_2) , then reject H_0 (i.e. that there is no statistical evidence that our null hypothesis is true at this confidence level)

Major Statistical Testing for Least Squares Estimation Problems

1. Test on the a posteriori variance factor.
2. Test on the goodness-of-fit for the estimated residuals.
3. Tests for outlying residuals.

1. Test on the *a posteriori* variance factor

- Although the least squares estimates $(\boldsymbol{v}, \hat{\boldsymbol{l}},\ \boldsymbol{and}\ \hat{\boldsymbol{x}})$ are independent of the choice of the *a priori* variance factor, their covariance matrices are not.
- Thus, testing the *a posteriori* variance factor is required for realistic covariance matrix elements.
- In fact, **the test on the *a posteriori* variance factor** is the **first step** in post-adjustment statistical testing (as it indicates in a global sense the presence of blunders or un-modelled systematic errors or improper functional model). **Later tests depend on its outcome** as well.
- Essentially, the test compares the *a priori* variance factor with the *a posteriori* variance factor. If the *a priori* variance factor was chosen correctly, then the a posteriori variance factor should be close to it.
  - So, the null hypothesis H_0 is: $H_{0:}: \dfrac{\hat{\sigma}_0^2}{\sigma_0^2} = 1$
  - This can be tested using the test statistic *y*: $y = \dfrac{r\hat{\sigma}_0^2}{\sigma_0^2}$ where

$$\hat{\sigma}_0^2 = \frac{v^T P v}{r}$$

where r is the adjustment's redundancy (degrees of freedom).

- This test statistic has a χ^2 distribution with $\boldsymbol{r}$ degrees-of-freedom. Consequently, the two tailed test at a significance level of $\boldsymbol{\alpha}$ is given by

$$\chi^2_{r,\alpha/2} \le y \le \chi^2_{r,1-\alpha/2}$$

- If the test fails, then we have grounds to believe that:
  1. σ_0^2 was chosen incorrectly.
  2. **The presence of outliers** (blunders) in the observations and therefore the residual vector (v) will contain larger values (**check the observations for outliers using another test**).
  3. **Weights are too large** (i.e. variances are small which means we trust the observations more than we should – **to remedy this we usually increase the observation variance-covariance matrix**).
  4. The model relating the observations and parameters was incorrect, or there were systematic errors not properly modelled.

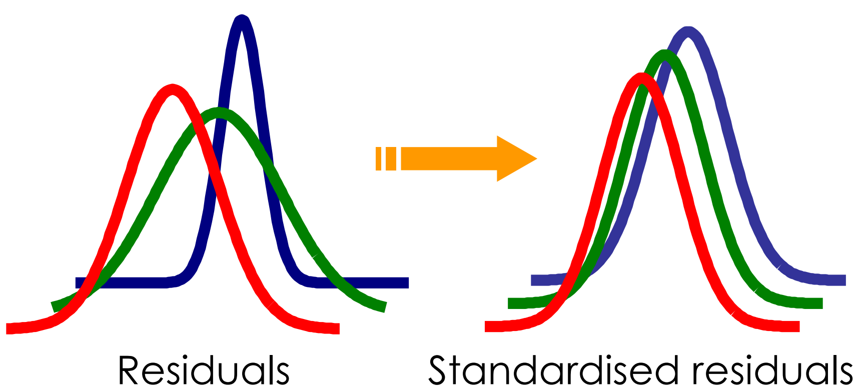
2. χ^2 goodness of fit test

- Adjustment using least-squares does not require that the observations be normally distributed (i.e. knowledge of the PDF of the observations is not required for least squares)
- However, if the observations are normally distributed then least squares will give "optimal" results (in a statistical sense).
- Also, statistical testing of adjustment results presupposes that the observations were normally distributed.
- The χ^2 goodness-of-fit test is done to check if the observations were, indeed, normally distributed.
- The test is done using the residuals, which will also be normally distributed.

Steps:

1. Calculate the estimated residuals and their standard deviations.
2. Each residual has an expected value of zero, but a different variance, and, consequently, a different normal distribution. To transform the residuals so that they all have the same distribution, "standardize" them by dividing each residual by its standard deviation.

$$\hat{v}_i^* = \frac{\hat{v}_i}{\sigma_{\hat{v}_i}}$$

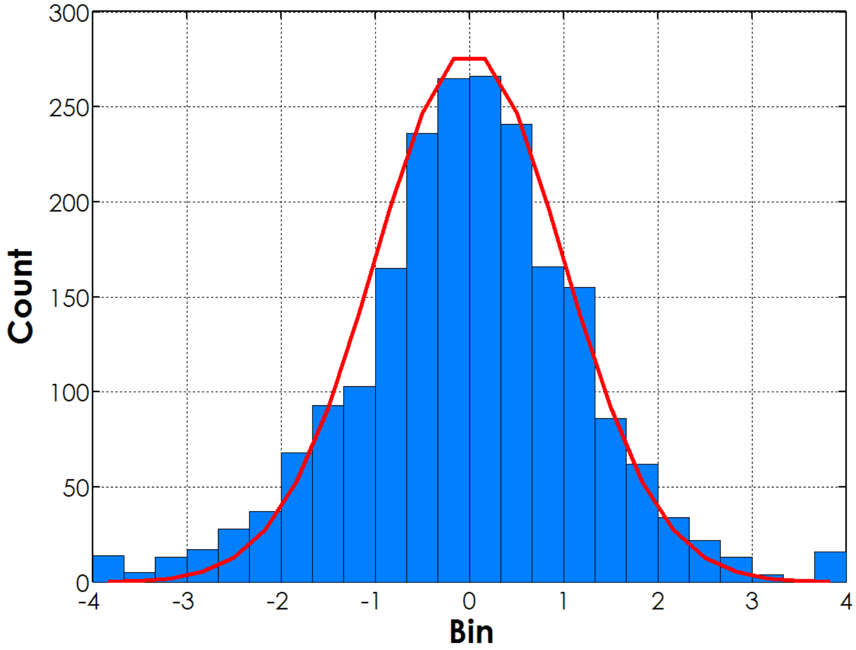


3. Bin the standardized residuals into evenly divided categories. Because the residuals have been standardized, they typically will be limited to values of ±3 (approximate 3-sigma). Typical bin sizes are 1/3 to 1/2.
4. For each bin, compute the expected number of residuals using the normal distribution, and the chi-squared value, given by

$$\frac{(a_i - e_i)^2}{e_i}$$

where a_i is the actual number of standardized residuals in bin (i) , e_i is the theoretical number from the normal distribution, and (i) is the bin number. e_i is simply the area under the standard normal curve between the upper and lower limit of bin *(i) multiplied by the total number of residuals:*

$\boldsymbol{e_i = p_i}$ **x n** , where
p_i is the population frequency from the area under the standard normal curve



5. The test statistic is the sum of the chi-squared values over all ***m*** bins,

$$y = \sum_{i=1}^{m} \frac{(a_i - e_i)^2}{e_i}$$

This test statistic has a χ^2 distribution with *m-1* degrees-of-freedom. Consequently, the one-tailed test at a significance level of α is given by:

$$0 \le y \le \chi^2_{m-1,\alpha}$$

If the test fails, then we have grounds to believe that:
1. The observations are not normally distributed.
2. There are outlying observations.

3. Test for outlying observations

- An outlier is an observation that contains a blunder
- Outlying observations can "silently" impact adjustment results (recall from Chapter 1 that outliers should be detected and removed from the observations before using them in the adjustment computation).
- In other words, the adjustment can report results and covariances that are not consistent with each other.
- The detection of outlying residuals is called **"data-snooping".** It is done using the adjusted residuals and their standard deviations.
- Each standardized residual is expected to belong to the standard normal distribution n(0, 1)

    - The null hypothesis H_0 is, then, $$H_0 : \hat{v}_i^* \in n(0,1)$$

  where, as before, $\hat{v}_i^*$ is given as : $$\hat{v}_i^* = \frac{\hat{v}_i}{\sigma_{\hat{v}_i}}$$

- H_0 is accepted at a significance level of α if

$$-Z_{\alpha/2} \leq \hat{v}_i^* \leq Z_{\alpha/2}$$

- If the test fails, then the residual is flagged as an outlier. **Typically, the adjustment should be repeated with the corresponding observation eliminated.**
- However, due to the smoothing effect of least squares, **residuals flagged as outliers may not correspond directly to the outlying observations**. Care should be taken when eliminating observations.

Examples on Hypothesis Testing:

1. Test on hypothesis for the population mean:

We have two cases:

a) μ (known ….. to be tested) , σ^2 known given $\bar{X}$ and n of a sample

$$\therefore y = \frac{\bar{X} - \mu}{\sigma / \sqrt{n}}$$ which is standard normal distribution

b) μ (known ….to be tested), σ^2 unknown given $\bar{X}$, S^2 and n of a sample

$$\therefore y = \frac{\bar{X} - \mu}{S / \sqrt{n}}$$ which is a t-distribution with (n-1) degrees of freedom

Example: A baseline of calibrated length (μ) of 400.008 m is measured 20 (n) times. The average of the measurements $(\bar{X})$ is 400.012 m with a standard deviation (s) of 0.002 m. Is the measured distance significantly different from the calibrated distance at a 0.05 (α) level of significance?

Steps:

1. H_0: μ = 400.012 H_a: $\mu \neq$ 400.012

$$y = \frac{\bar{X} - M}{S\sqrt{n}} = \frac{400.012 - 400.008}{0.002/\sqrt{20}} = 8.944$$

1. y has a t-distribution test (falls under the b) case)
2. Due to symmetry of the t-distribution

H_0 is rejected if: $|y| > t_{\alpha/2,(n-1)}$ where $t_{0.025,19} = 2.093$

3. Make a decision

$$|y| = 8.944 > t_{0.025,19} = 2.093 \rightarrow \text{reject } H_0$$

i.e. there is a reason to believe that the average measured distance is significantly different from the calibrated distance at 5% significance level

Note: If you want to check the result, you can also calculate the confidence interval on μ:

$$\left[t_{1-\alpha/2,19} < \frac{\overline{X}-\mu}{S/\sqrt{n}} < t_{\alpha/2,19} \right] = 95\%$$

$$\left[\overline{X} - t_{1-\alpha/2,19}\frac{S}{\sqrt{n}} < \mu < \overline{X} + t_{\alpha/2,19}\frac{S}{\sqrt{n}} \right] = 95\%$$

$$\left[400.012 - 2.093\frac{0.002}{\sqrt{20}} < \mu < 400.012 + 2.093\frac{0.002}{\sqrt{20}} \right] = 95\%$$

$$\left[400.011 < \underset{400.008}{\overset{\mu}{\downarrow}} < 400.013 \right] = 95\%$$

does not lie between the specific interval

Example: The manager of a surveying firm wants all surveying technicians to be able to read a particular instrument to within ± 1.5”. To test this value the senior field crew chief is asked to perform a reading test with the instrument. The crew chief reads the circle 30 times and obtains a S_r=±0.9”. Does this support the 1.5” limit at a 5% significance level?

Solution: In this case, the manager wishes to test the hypothesis that the computed sample variance is the same as the population variance, versus its being greater than the population variance. That is, all standard deviations that are equal to or less than 1.5” will be accepted. Thus a one-tailed test is constructed as follows. (note that m = 30 – 1,or 29.)

The null hypothesis is: $H_{0:} : S^2 \le \sigma^2$

The alternative hypothesis is: $H_a : S^2 > \sigma^2$

The test statistic is: $y = \frac{r\,S^2}{\sigma^2} = \frac{(30-1)(0.9)^2}{1.5^2} = 10.44$

and the null hypothesis is rejected when the computed test statistic exceeds the tabulated value, or when

$$y > X^2_{\alpha,v}$$

In this case $X^2_{0.05,29}$ = 42.56, and since the computed test statistic value (10.44) is less than the tabulated value (42.56), the null hypothesis is not rejected. However, simply failing to reject the null hypothesis does not mean that the value of ± 1.5” is valid.

Example: Ron and Kathi continually debate who measures angles more precisely with a particular instrument. Their supervisor, after hearing enough, describes a test where each is to measure a particular direction by pointing and reading the instrument 51 times. They must then compute the variance for their data. At the end of the 51 readings, Kathi determines her variance to be 0.81, and Ron finds his to be 1.21. Is Kathi a better instrument operator at a 0.01 level of significance?

Solution: In this situation, even though Kathi's variance implies that her measurements are more precise than Ron's, a determination must be made to see if the reference variances are statistically equal, versus Kathi's being better than the Ron's. This test requires a one-tailed F-test with a significance level of α = 0.01.

The null hypothesis is: $H_0: \frac{S_R^2}{S_K^2} = 1 \quad \left(S_R^2 = S_k^2\right)$

And the alternative hypothesis is: $H_a: \frac{S_R^2}{S_K^2} > 1 \quad \left(S_R^2 > S_K^2\right)$

The test statistic is: $\mathrm{y} = \frac{1.21}{0.81} = 1.49$

Again, the null hypothesis is rejected when the computed test statistic exceeds the tabulated value:

$$\mathrm{y} > \mathrm{F}_{\alpha,\mathrm{v1},\mathrm{v2}}$$

In this case, $\mathrm{F}_{0.01,50,50} = 1.95$, and since the computed test statistic value (1.49) is less than the tabulated value (1.95), the hull hypothesis is not rejected. Therefore, there is no statistical reason to believe that Kathi is better than Ron at a 0.01 level of significance.

Example: A baseline is repeatedly measured using a EDM instrument over a five-day period. Each day, 10 measurements are taken and averaged. The variances for the measurements are listed below. At a significance level of 0.05, are the results of day 2 significantly different from those of day 5?

Day	1	2	3	4	5
Variance, S^2 (mm^2)	50.0	61.0	51.0	53.0	54.0

Solution: This problem involves checking whether the variances of day 2 and day 5 are statistically equal versus them being different. This is the same as constructing a confidence interval involving the ratio of the variances. Because the concern is about equality or inequality, this will require a two-tailed test. Since there were 10 values collected each day, both variances are based on 9 degrees of freedom (v_1 and v_2). Assume that the variance for day 2 is S_2^2 and the variance for day 5 is S_5^2. The test is constructed as follows:

The null hypothesis is $H_0: \frac{S_2^2}{S_5^2} = 1$

and the alternative hypothesis is $H_a: \frac{S_2^2}{S_5^2} \neq 1$

The computed test statistic is $\mathrm{y} = \frac{61}{54} = 1.13$

Again, the null hypothesis is rejected when the computed test statistic exceeds the tabulated value:

$$\mathrm{y} > \mathrm{F}_{\alpha/2,\mathrm{v1},\mathrm{v2}}$$

In this case, $\mathrm{F}_{0.025,9,9} = 4.03$, and since the computed test statistic (1.13) is less than the tabulated value (4.03), the null hypothesis is not rejected. Consequently, there is no statistical reason to believe that the data of day 2 are statistically different from those of day 5.