Lecture Notes on Statistical Methods

(by Tom Co 9/23/2007, 10/15/2007)

Charateristics of a Good Engineering Experiment

1. Necessity.
- a) objective is well formulated
- b) economical
- c) results are needed for decision, understanding and process improvement

2. Scope.
- a) significant variables are tested within important range
- b) (boundary and initial) conditions are properly set up
- c) results are representative of general case, e.g. scalable

3. Reproducibility and Statistical Significance
- a) enough trials need to be taken to assess confidence
- b) results must be reproducible for accuracy and precision of prediction

4. Realization
- a) results can be applied to real process or system
- b) data are relevant to the real problem

5. Analysis
- a) statistical analysis of data can and are applied
- b) the quality and confidence of the results including models are properly assessed

General Concepts

1. Random Variable

- a measured variable that takes on a range of possible values which are random (i.e. lacking exact predictability)

Two types of random variables:

a. discrete

Example: N_R = number of ceramic rasching rings per cubic feet of absorption column

b. continuous

Example: V_f = the void fraction per cross section area of the absorption column

2. (Statictical) Event

- an occurence of the random variable taking on some specified values or range of values.

Example: the number of ceramic rasching rings per cubic feet is greater than 200

$$(N_R > 200)$$

Example: the void fraction per unit per cross section area is between 0.25 and 0.5

$$(0.25 \leq V_f \leq 0.5)$$

3. Probability

- The likelihood (normalized frequency) for the occurence of an event.

  Example: $\Pr\,(0.25 \leq V_f \leq 0.5) = 0.25$

  Special case: When random variable is discrete, then discrete probability is the ratio of the [number of cases favorable to an event] to the [number of all possible cases] also known as the frequency of the event.

(For a list of properties of probabilities, see Appendix 1.)

4. Probability Distribution

- a function (or mapping) of events to probabilities

Motivation:

Using historical data and experience (or assumptions), we want a convenient way to estimate or predict probabilities of events

Methods:

a. <u>Using histograms</u>

- a grouping of collected data into categorized bins (e.g.\ range of values)

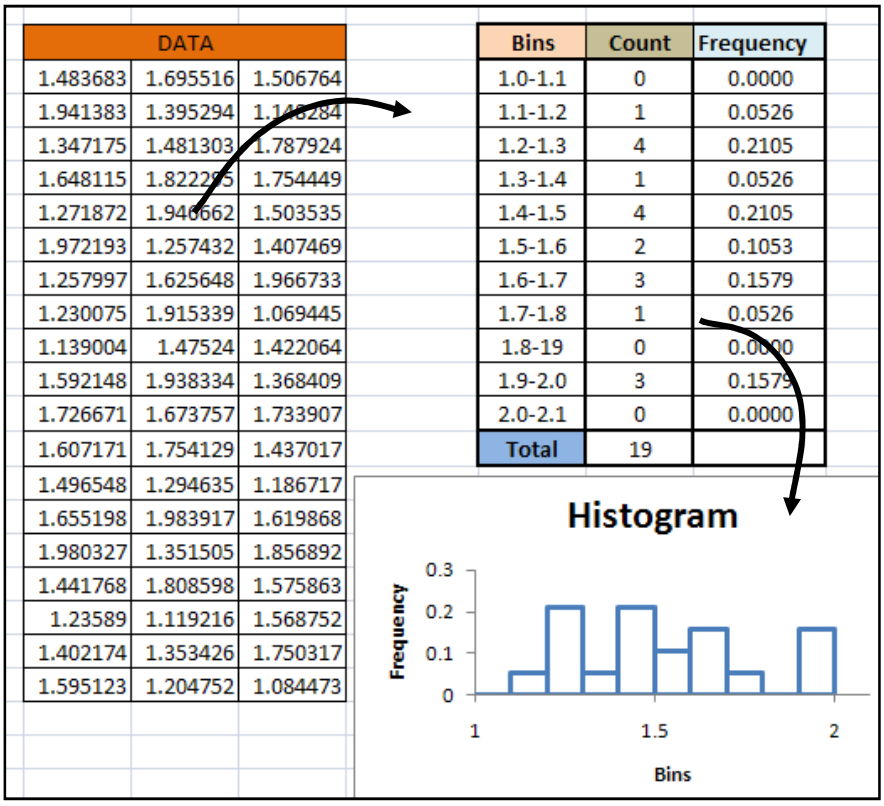


Figure 1.

$$\Pr(1.4 \le x \le 1.8) = 0.2105 + 0.1053 + 0.1579 + 0.0526 = 0.5263$$

(See Appendix 7 for details on using Excel to create histograms.)

b. Using probability density functions (pdf)

- a continuous approximation of a frequency histogram

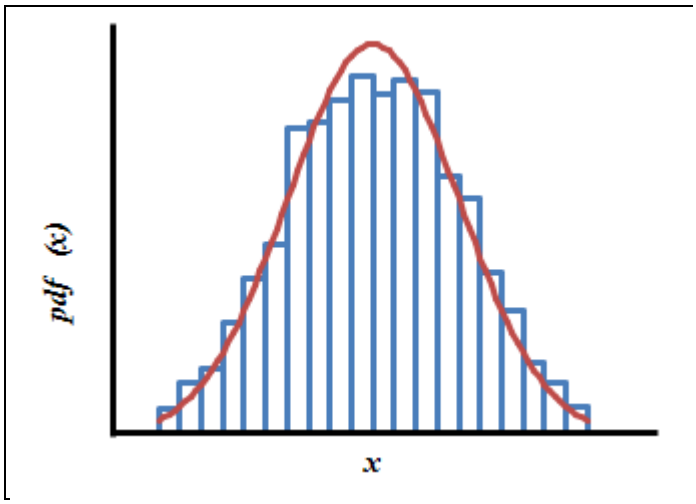


Figure 2.

$$\Pr(a \le x \le b) = \int_a^b pdf(x)dx$$

For a list of important probability distributions, see Appendix 2 and 3.

- for discrete random variables, the function becomes the *probability mass function* (pmf) which has relevance only at the discrete points.

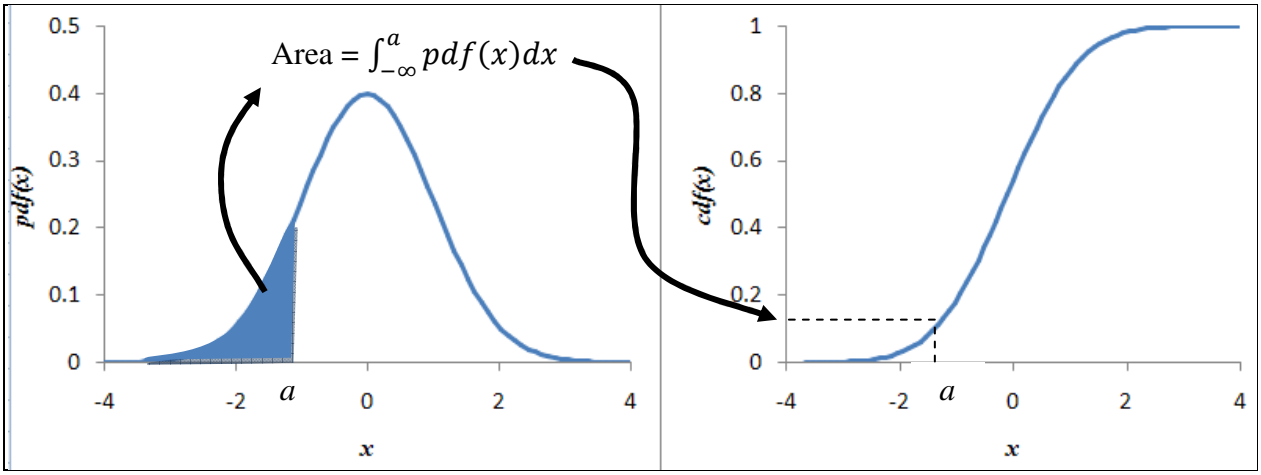
$$\Pr(\ x = a\ ) = pmf(a)$$

(Examples of these are given in Appendix 2.) They are usually represented as a curve with dots at the discrete points; or, if the discrete random variables are spread evenly, the pmf can be represented by bar-charts.

c. Using cumulative distribution functions (cdf)

- a distribution that yields the probabilities of a one-sided range of random variables

$$cdf(a) = \Pr\ (x \le a) = \int_{-\infty}^{a} pdf(x)dx$$



- For discrete random variables, the cumulative probabilities are given by

$$\Pr(x \leq a) = \sum_{x \leq a} pmf(x)$$

Measures of Central Tendency:

Let $f(x)$ and $m(x)$ be the probability density function and probability mass function, respectively, of the population:

a) Population Mean ("Expected Value of x")

$$\mu = \int_{-\infty}^{\infty} x\, f(x)dx \quad for\ \ continuous\ random\ variables$$

or

$$\mu = \sum_{-\infty \leq x \leq \infty} x\, m(x) \quad for\ discrete\ random\ variables$$

b) Sample Mean (Average)

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

Measures of Variability

a) Population Variance:

$$\sigma^2 = \int_{-\infty}^{\infty} (x-\mu)^2 f(x)dx \quad for\ continuous\ random\ variables$$

or

$$\sigma^2 = \sum_{-\infty \le x \le \infty} (x-\mu)^2 \, m(x) \quad for\ discrete\ random\ variables$$

b) Sample Variance:

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

The population standard deviation and sample standard deviation are given by σ and s, respectively.

Other Measures:

i. Median: 50% of the population is less than the median point

$$\Pr(-\infty \le x \le x_{median}) = 0.5$$

ii. The first quartile (25th percentile) and third quartile (75th percentile) can be used to identify outliers (see appendix 6 for details).

iii. Mode: peak points of the probability distribution functions,

$$\frac{df}{dx} = 0 \quad and \quad \frac{d^2 f}{dx^2} < 0$$

Some Important Properties:

1. The binomial distribution has: mean: $\mu = np$ and variance: $\sigma^2 = np(1-p)$.
2. As n becomes large, the binomial distribution approaches a normal distribution
3. The mean of a normal distribution is μ while the standard deviation is σ.
4. Define a new variable z, known as the standard scores, as

$$z = \frac{x-\mu}{\sigma}$$

If x is normally distributed with mean μ and standard deviation σ,
z will follow a standard normal distribution with mean equal to zero and standard deviation equal to one.

5. Let $x_1, x_2, \ldots, x_n$ be n samples taken independently from the same population with a fixed probability distribution, then the sum

$$W = \sum_{i=1}^{n} x_i$$

will approach a normal distribution as n approaches infinity.

6. In particular, the sample average, i.e. $\bar{x} = W/n$, will be normally distributed with a mean equal to that of the original population. This is also known as the Central limit theorem.

7. Another result of the Central limit theorem is that the standard deviation of the distribution of the sample averages will be equal to $(\sigma/\sqrt{n})$. (See Appendix 5 for derivation of this fact.)

8. If n is small (e.g. <20), a correction to the Central limit theorem is to use a t-distribution instead, with degree of freedom, v=(n-1), where the t-scores are used instead of z-scores

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}}$$

9. Let Y_1, Y_2, …, Y_N be independent N random variables, each following a standard normal distribution. Then the sum of squares given by

$$Q = \sum_{i=1}^{N} Y_i^2$$

will follow a Chi-square distribution with the degree of freedom, $\nu = N$. (For every constraint imposed on the N random variables, the degree of freedom is reduced accordingly. For instance, if the sum of random variables has to be equal to a fixed number, say 120, then the degree of freedom is reduced by 1.)

Application 1: Generating Confidence Intervals for Sample Means

Main Problem:

- The sample mean $\bar{x}$ is supposed to estimate the population mean μ. This will yield only a "point"-estimate, which has a very low probability of being exactly equal to μ.
- Instead, we want to generate an interval, e.g. $(\bar{x} - \epsilon, \bar{x} + \epsilon)$ such that we are confident, within a prescribed confidence level, that the real value of μ is inside this interval.

Procedure:

Example: Consider a 10-sample set given by

1.234	1.209	1.213	1.231	1.223
1.225	1.23	1.22	1.218	1.216

1. Determine the value of t-score that would yield the required confidence level based on t-distribution.

   Example: For a confidence level of 95%, we want to find the value of t from the t-distribution in which the two-tail probability is equal to 5%. Since $n = 10$, the degree of freedom is 9. Using the Excel function, we find

   $$t_{\text{confidence interval}} = \text{TINV}(0.05,9) = 2.2622$$

2. Calculate sample average $\bar{x}$ and sample standard deviation s.

   Example: (from above data)

   $$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = 1.2219, \qquad s = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}} = 0.008198$$

3. Calculate the interval estimate based on t-scores : (Note: the value $s/\sqrt{n}$ is also known as the "standard error".)

   $$t = \frac{\bar{x} - \mu}{s/\sqrt{n}} \quad \Longrightarrow \quad \epsilon = (\bar{x} - \mu) = t_{\text{confidence interval}} \frac{s}{\sqrt{n}}$$

   Example: (continuing from above)

   $$\epsilon = 2.2622 \frac{0.008198}{\sqrt{10}} = 0.005865$$

   Thus, with 95% confidence, the population mean can be estimated as

   $$\mu \in (1.2219 - 0.005865,\ 1.2219 + 0.005865) = (1.2160,\ 1.2278)$$

Application 2: Calculating Sample Sizes

Main Problem:

- We want to estimate the population mean μ to within a specified precision.
- Assuming we have a reasonable idea of the standard deviation σ of the population, we need to determine how many samples are needed in order to satisfy the required precision.

Procedure:

Example: The recipe of a batch process is known to yield products that have a standard deviation of 0.02 g/liter of impurities.

We want to determine how many batch samples to measure such that within a 95% confidence interval, the measured concentration of impurities will be ±0.015 g/liter of the sample mean.

1. Since we have not yet done the actual measurements, we assume that the sample standard deviation s is the same as the population standard deviation σ. This assumption will allow us to calculate the critical values based on the t-distribution. For a 95% confidence interval,

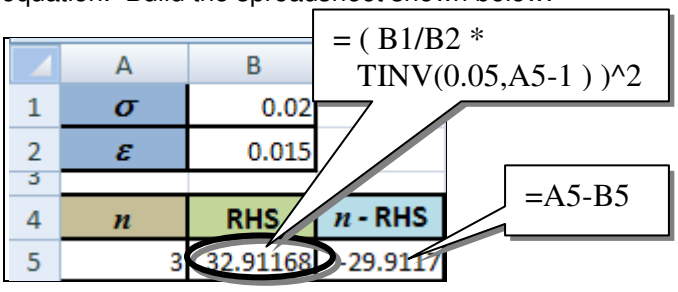
$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}} = \frac{\epsilon}{\sigma/\sqrt{n}}$$

$$\Rightarrow \quad n = \left(t_{\text{confidence interval}} \frac{\sigma}{\epsilon}\right)^2 = \left(\frac{\sigma}{\epsilon}\right)^2 [\text{TINV}(0.05, n-1)]^2$$

Example: Based on the given values above, i.e. $\sigma = 0.02$ g/liter and $\epsilon = 0.015$ g/liter,

$$n = \left(\frac{0.020}{0.015}\right)^2 [\text{TINV}(0.05, n-1)]^2 \qquad \text{(Eqn 1)}$$

Note that n appears on both sides of the equation. One approach is assume a standard normal distribution instead of the t-distribution. However, this is valid only if the sample size is large. A more accurate method is to use computational tools such as an Excel spreadsheet to solve the equation. Build the spreadsheet shown below:



Where RHS stands for the right hand side of equation (1). Then obtain the smallest value of n (must be integer) such that the value of n-RHS is positive. As shown below, we need n=10 batch samples.

n	RHS	n - RHS
10	9.09752	0.90248

Application 3: Hypothesis Testing : Whether Two Sample Means are Significantly Different

Main Problem:

- Given two sample groups of size n_1 and n_2, each yielding sample means $\bar{x}_1$ and $\bar{x}_2$, and standard deviations, s_1 and s_2.
- Assuming both samples are obtained from the same populations with the same standard deviation σ, we want to determine whether the two sample means are significantly different (based on a desired confidence level.)

Procedure:

Example: Two sets of samples of the distillate concentrations were collected one week apart, yielding the following calculations:

	n	$\bar{x}$	s
Week 1	10	0.922	0.012
Week 2	20	0.903	0.015

We want to know if the mean of week 2 is significantly different from the mean of week 1 using a 99% confidence interval.

1. Set up the null hypothesis $\mathbf{H_o}$ and the alternative hypothesis $\mathbf{H_a}$.

$\mathbf{H_o}$	$\bar{x}_1 = \bar{x}_2$
$\mathbf{H_a}$	$\bar{x}_1 \neq \bar{x}_2$

2. Calculate the critical value of *t*-distribution needed for a confidence interval of 99% confidence interval, with a degree of freedom equal to : $n_1 + n_2 - 2$.

   Example: From our given values, the degree of freedom is 20+10-2=28. With a 99% confidence interval, we find:

$$t_{\text{critical}} = \text{TINV}(0.01,28) = 2.763$$

3. Next, calculate a pooled standard deviation given by the formula:

$$s_p = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}}$$

   Example: From our given values, we find:

$$s_p = \sqrt{\frac{(9)0.012^2 + (19)0.015^2}{28}} = 0.01410$$

4. Calculate the t-score for the difference of the sample means

$$t = \frac{\bar{x}_1 - \bar{x}_2}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

Example: From our given values, and the calculated s_p:

$$t = \frac{0.922 - 0.903}{0.01410\sqrt{\frac{1}{10} + \frac{1}{20}}} = 3.478$$

5. Compare t score with t_{critical}. If the t score is beyond the confidence interval, then we reject the null hypothesis and accept the alternative hypothesis.

   Example: Since $t > t_{\text{critical}}$, i.e. 3.478>2.763, we reject the null hypothesis and conclude that the sample mean of week 2 is significantly different from the sample mean of week 1.

Remarks:

a. There are two types of errors that are possible when using hypothesis testing. Type 1 error is the error when the null hypothesis was true but was rejected. Type 2 error occurs when the null hypothesis was false but was accepted.

b. The hypothesis testing method can be used in several other comparisons. Appendix 4 lists some of the important cases together with the type of distributions used to determine confidence intervals.

c. The table in Appendix 4 shows two entries when comparing two sampled means. Entry 8 considers the case in which the standard deviations of the populations used for each sample group are the same (this was discussed in this section). Entry 9, however, considers the case when the standard deviation may have been different for the population that yielded the sample means. In this case, the degree of freedom requires a more complicated evaluation that may yield a non-integer result. When this happens, the fractional part is simply dropped.

Appendix 1. Properties of Probabilities

Let A and B be events. $\Pr(A|B)$ is the conditional probability, i.e. the probability of event A on the condition that event B has occurred.

1. General relationships:

1	$\Pr(\text{ all events }) = 1$ $\Pr(\text{ no events }) = 0$
2	$\Pr(\text{ not } A\ ) = 1 - \Pr(A)$
3	$\Pr(A \cap B) = \Pr(A\|B)\Pr(B) = \Pr(B\|A)\Pr(A)$
4	$\Pr(A \cup B) = \Pr(A) + \Pr(B) - \Pr(A \cap B)$

2. Special Cases:

1	Events A and B are independent	$\Pr(A\|B\ ) = \Pr(A)$ $\Pr(B\|A) = \Pr(B)$ $\Pr(A \cap B) = \Pr(A)\Pr(B)$
2	Events A and B are mutually exclusive, i.e. $A \cap B = 0$	$\Pr(\ A \cup B\ ) = \Pr(A) + \Pr(B)$

3. Bayes' Formula:

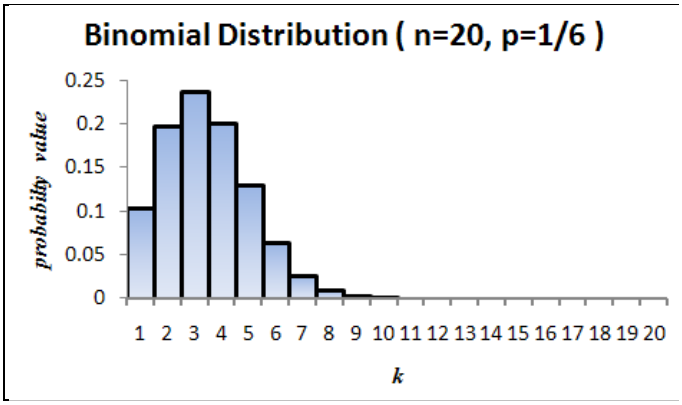
$$\Pr(B|A) = \frac{\Pr(A|B)\Pr(B)}{\Pr(A)}$$

Appendix 2. Some Discrete Probability Distributions

1. Binomial Distribution

Let n be the number of independent trials, k be the number of successful occurrence and p be the probability of success for a single trial, then

$$\Pr(x = k \mid n, p) = \frac{n!}{k!\,(n-k)!}\, p^k (1-p)^{1-k}$$



Available Excel functions:

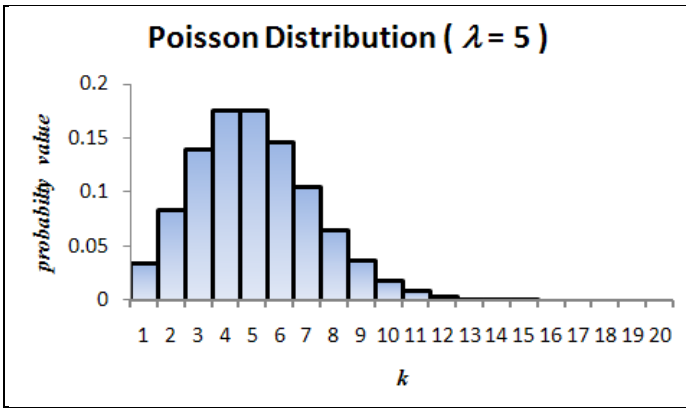
BINOMDIST(k, n, p, cum)	Binomial probability distribution function $(0 \le k \le n)$	k = number of success n = number of trials p = probability of single trial cum = TRUE (for cumulative) and FALSE (for probability)

2. Poisson Distribution

Let k be the number of successful occurrence in τ time units, λ be the expected number of successful occurrences, i.e. let $\bar{t}$ be the average time for a successful occurrence, then

$$\lambda = \frac{\tau}{\bar{t}}$$

$$\Pr(x = k \mid \lambda) = \frac{\lambda^k}{k!} e^{-\lambda}$$



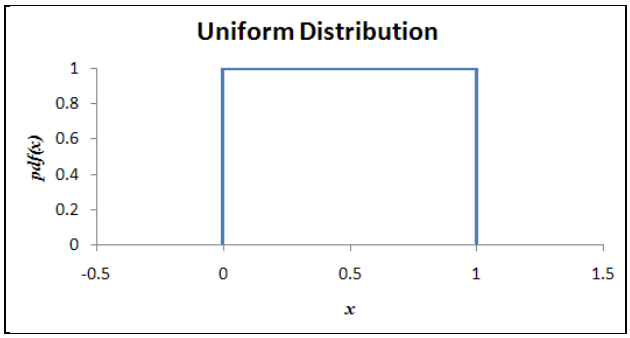
Available Excel function:

POISSON(k, λ, cum)	Poisson probability distribution function ($0 \leq k, \lambda$)	k = number of success λ = expected number of success cum = TRUE (for cumulative) and FALSE (for probability)

Appendix 3. Some Continuous Probability Distributions

1. Uniform Distribtuion

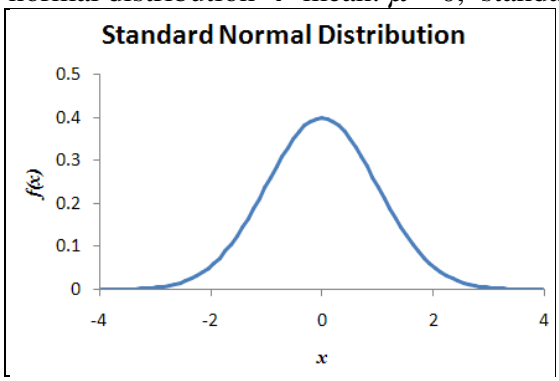
$$f(x) = 1 \quad where \quad 0 \le x \le 1$$



2. Normal Distribution (also known as Gaussian distribution and denoted $N(\mu,\sigma)$)

$$f(x) = \frac{1}{\sigma\sqrt{2\pi p}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

Special Case: standard normal distribution → mean: $\mu = 0$, standard deviation : $\sigma = 1$



Remarks:

a. The normal distribution function is symmetric around the mean μ.

b. Available functions in Excel are:

NORMDIST(x, μ, σ, cum)	Normal distribution (if cum=FALSE, then yield normal pdf)	where, x = random variable z = standard score $= \frac{x-\mu}{\sigma}$ μ = mean σ = standard deviation Pr = probability
NORMSDIST(z)	Standard normal cumulative distribution	
NORMINV(Pr, μ, σ)	Inverse normal cumulative distribution	
NORMSINV(Pr)	Inverse standard normal cumulative distribution	

Example 1: Suppose the random variable x is known to be normally distributed with a mean of 3 and standard deviation of 0.6. Determine $\Pr(2 \leq x \leq 5)$.

Solution:

NORMDIST(5, 3, 0.6, TRUE)-NORMDIST(2, 3, 0.6, TRUE) =0.952

Or with $z_1 = \frac{5-3}{0.6} = 3.3333$ and $z_2 = \frac{2-3}{0.6} = -1.6667$
NORMSDIST(3.3333)-NORMSDIST(-1.6667) = 0.952

Example 2: Suppose the random variable x is known to be normally distributed, determine the mean and standard deviation such that $\Pr(x \leq 7) = 0.3$ and $\Pr(x \leq 10) = 0.8$.

Solution:

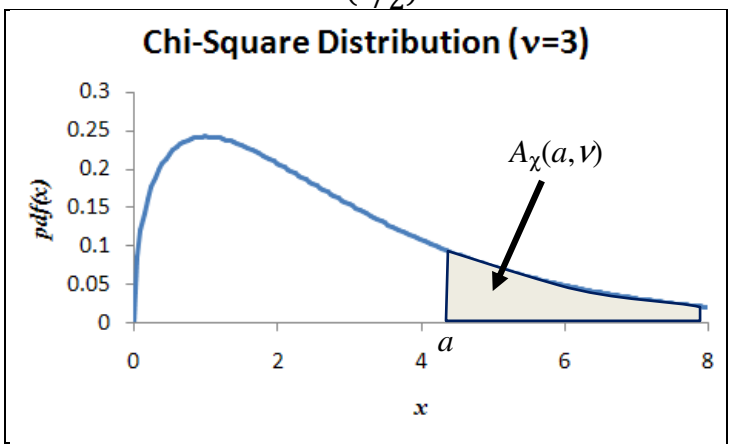
z_1 = NORMSINV(0.8) = 0.8416 = $\frac{10-\mu}{\sigma}$; z_2 = NORMSINV(0.3) = -0.5244 = $\frac{7-\mu}{\sigma}$

Solving simultaneously for mean and standard deviation, we get

$$\begin{pmatrix} 1 & 0.8416 \\ 1 & -0.5244 \end{pmatrix} \begin{pmatrix} \mu \\ \sigma \end{pmatrix} = \begin{pmatrix} 10 \\ 7 \end{pmatrix} \quad \rightarrow \quad \mu = 8.1517 \text{ , } \sigma = 2.1962$$

3. <u>Chi-Square (χ^2) Distribution</u>

$$f(x) = \left(\frac{1}{2}\right)^{\nu/2} \frac{1}{\Gamma(\nu/2)} x^{(\upsilon/2-1)} e^{-\nu/2}$$



Define the right tail area by:

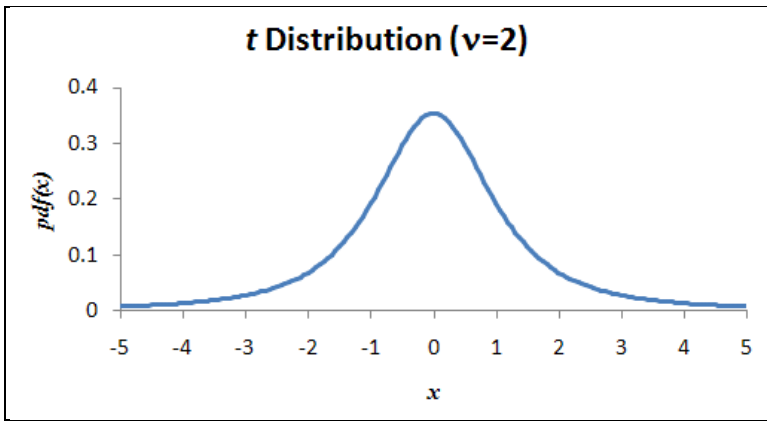
$$A_\chi(a,\nu) = \int_a^\infty f_\chi(x;\nu)\, dx \qquad \text{where } f_\chi(x;\nu) \text{ is the } \chi^2 \text{ pdf}$$

Available Excel functions:

CHIDIST(x, ν)	Right-tail probability of a Chi-square distribution	where, x = random variable ν = degree of freedom
CHIINV(Pr, ν)	Inverse of the right-tail probability of a Chi-square distribution	

4. <u>t – Distribution</u>

$$f(x) = \frac{1}{\sqrt{\pi\nu}} \frac{\Gamma\left(\frac{(\nu+1)}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)} \left(1 + \frac{x^2}{\nu}\right)^{-\frac{(\nu+1)}{2}}$$



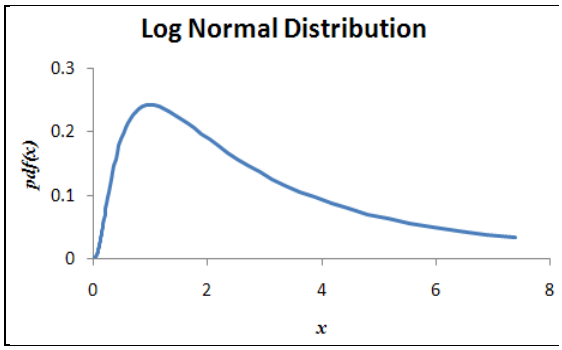
Available Excel functions:

TDIST(x, ν, 1)	Right-tail probability of a t distribution (x is nonnegative)	where, x = random variable ν = degree of freedom
TDIST(x, ν, 2)	Two-tail probability of a t distribution (x is nonnegative)	
TINV(Pr, ν)	Inverse of the two-tail probability of a t distribution	

Other Important Probability Distributions:

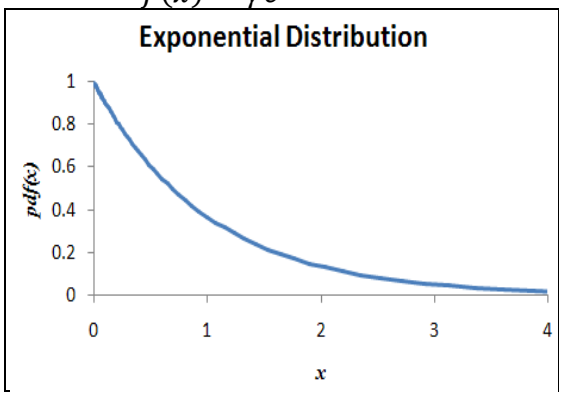
5. <u>Log Normal Distribution</u>

$$f(x) = \frac{1}{x\beta\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{\ln(x)-\alpha}{\beta}\right)^2}$$



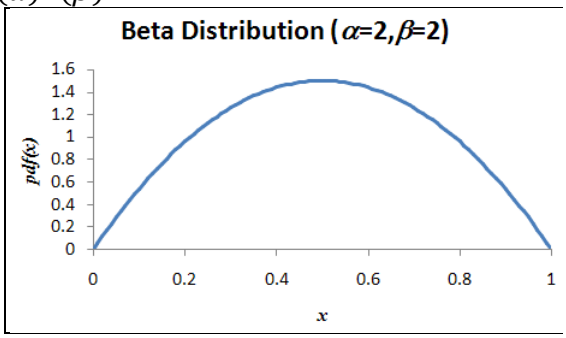
6. Exponential Distribution (Note: discrete version is a Poisson distribution)

$$f(x) = \gamma e^{-\gamma x}$$



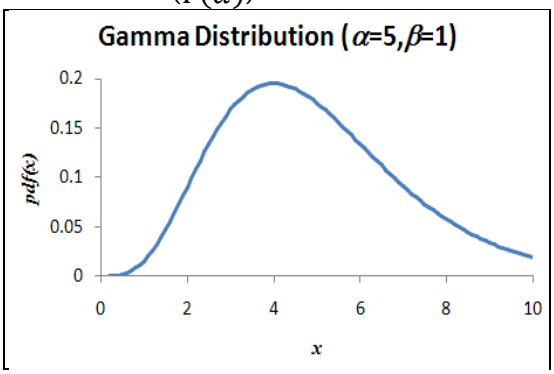
7. Beta Distribution

$$f(x) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1} \quad ; \quad \Gamma(\cdot) = \text{gamma function}$$



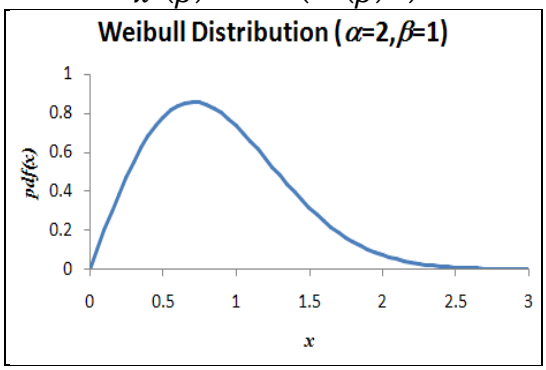
8. Gamma Distribution

$$f(x) = \left(\frac{\beta^{\alpha}}{\Gamma(\alpha)}\right) x^{\alpha-1} e^{-\beta x}$$



9. Weibull Distribution

$$f(x) = \frac{\alpha}{x}\left(\frac{x}{\beta}\right)^{\alpha} \exp\left(-\left(\frac{x}{\beta}\right)^{\alpha}\right)$$



Available functions in Excel:

LOGNORMDIST(x, α, β)	Log normal cumulative distribution (x is nonnegative)	where, x = random variable α, β, γ = parameters Pr = probability A = lower bound B = upper bound cum = TRUE(cdf) or FALSE(pdf)
LOGINV(Pr, α, β)	Inverse log normal cumulative distribution	
EXPONDIST(x, γ, cum)	Exponential distribution (x is nonnegative)	
BETADIST(x , α, β, A, B)	Beta cumulative distribution ($A \leq x \leq B$)	
BETAINV(Pr, α, β, A, B)	Inverse beta cumulative distribution	
GAMMADIST(x, α, β, cum)	Gamma distribution (x is nonnegative)	
GAMMAINV(Pr, α, β)	Inverse Gamma distribution	
WEIBULL(x, α, β, cum)	Weibull distribution	

Appendix 4. Table of Hypothesis Tests

	Test	Given/Calc				Statistic	Tails	Distribution	H_o
		μ	σ	$\bar{x}$	s				
1	x significantly different from μ	•	•			$\frac{x-\mu}{\sigma}$	2	$N(0,1)$	$x=\mu$
2	x significantly lower than μ						Left		$x \geq \mu$
3	x significantly higher than μ						Right		$x \leq \mu$
4	$\bar{x}$ significantly different from μ (σ known)	•	•	•		$\frac{\bar{x}-\mu}{\sigma}\sqrt{n}$	2	$N(0,1)$	$\bar{x}=\mu$
5	$\bar{x}$ significantly different from μ (σ unknown)	•		•	•	$\frac{\bar{x}-\mu}{s}\sqrt{n}$	2	t (n-1)	$\bar{x}=\mu$
6	Two sample means significantly different (σ same for both)		•	•		$\frac{\bar{x}_1-\bar{x}_2}{\sigma\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}$	2	$N(0,1)$	$\bar{x}_1=\bar{x}_2$
7	Mean for paired comparison of samples, $x_i=y_i-z_i$, is significantly nonzero			•	•	$\frac{\bar{x}}{s}\sqrt{n}$	2	t (n-1)	$\mu=0$
8	$\bar{x}_1 \neq \bar{x}_2$ (σ same for both)			•	•	$\frac{\bar{x}_1-\bar{x}_2}{S\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}$ with $S=q\,s_1+(1-q)s_2$ $q=\frac{n_1-1}{n_1+n_2-2}$	2	t (n_1+n_2-2)	$\bar{x}_1=\bar{x}_2$
9	$\bar{x}_1 \neq \bar{x}_2$			•	•	$\frac{\bar{x}_1-\bar{x}_2}{\sqrt{\frac{{s_1}^2}{n_1}+\frac{{s_2}^2}{n_2}}}$	2	$t\left(\frac{(\delta_1+\delta_2)^2}{\frac{{\delta_1}^2}{n_1-1}+\frac{{\delta_2}^2}{n_2-1}}\right)$ with $\delta_k=\frac{{s_k}^2}{n_k}$	$\bar{x}_1=\bar{x}_2$
10	${s_1}^2 > {s_2}^2$				•	${s_1}^2 \Big/ {s_2}^2$	Left	$F(n_1-1, n_2-1)$	${s_1}^2={s_2}^2$
11	$s \neq \sigma$				•	$(n-1)\frac{s^2}{\sigma^2}$	2	$\chi^2(n-1)$	$s=\sigma$

Appendix 5. Some Important Formulas for Means and Variances

1. Let $E(x)$ be the "expected value" of x, with respect to a probability distribution function given by $p(x)$, defined by the integral

$$E(x) = \int_{-\infty}^{\infty} x\, p(x)dx$$

2. The expected values of a sum of random variables is the sum of expected values:

$$\begin{aligned} E(x+y) &= \iint_{-\infty}^{\infty} (x+y)p(x)p(y)dydx \\ &= \left(\int_{-\infty}^{\infty} xp(x)dx \int_{-\infty}^{\infty} p(y)dy\right) + \left(\int_{-\infty}^{\infty} p(x)dx \int_{-\infty}^{\infty} yp(y)dy\right) \\ &= E(x) + E(y) \end{aligned}$$

3. The expected value of a product of independent random variable is the product of expected values:

$$\begin{aligned} E(xy) &= \iint_{-\infty}^{\infty} (xy)p(x)p(y)dydx \\ &= \left(\int_{-\infty}^{\infty} xp(x)dx\right)\left(\int_{-\infty}^{\infty} yp(y)dy\right) \\ &= E(x)E(y) \end{aligned}$$

4. The mean of the population, μ, is the expected value of random variable x,

$$\mu = E(x)$$

5. The variance of a random variable x is the expected value of $(x - E(x))^2$, i.e.

$$Var(x) = \sigma^2 = E\left(\left(x - E(x)\right)^2\right) = E((x-\mu)^2) = \int_{-\infty}^{\infty} (x-\mu)^2\, p(x)dx$$

6. The variance of a sum is the sum of variances:

$$\begin{aligned} Var(x+y) &= E\left(\left((x+y) - E(x+y)\right)^2\right) = E\left(\left[(x-\mu_x) + (y-\mu_y)\right]^2\right) \\ &= E\left((x-\mu_x)^2 + 2(x-\mu_x)(y-\mu_y) + (y-\mu_y)^2\right) \\ &= E((x-\mu_x)^2) + 2E(x-\mu_x)E(y-\mu_y) + E\left((y-\mu_y)^2\right) \\ &= Var(x) + Var(y) \end{aligned}$$

7. Variance of a scaled random variable, *kx* is $k^2Var(x)$:

$$Var(kx) = E\left(\left(kx - E(kx)\right)^2\right) = E\left(\left(kx - kE(x)\right)^2\right) = E((kx - k\mu)^2)$$
$$= E(k^2(x-\mu)^2) = k^2E((x-\mu)^2) = k^2Var(x)$$

8. The expected value of the sample mean is the population mean:

$$E(\bar{x}) = E\left(\frac{\sum_{i=1}^{n} x_i}{n}\right) = \frac{1}{n}\sum_{i=1}^{n} E(x_i) = \frac{1}{n}\sum_{i=1}^{n} \mu = \frac{1}{n}(n\mu) = \mu$$

Thus, the sample mean is an "unbiased" estimator of the population mean.

9. The variance of sample means is $\left(\sigma^2/n\right)$:

$$Var(\bar{x}) = Var\left(\frac{\sum_{i=1}^{n} x_i}{n}\right) = Var\left(\sum_{i=1}^{n}\frac{x_i}{n}\right) = \sum_{i=1}^{n} Var\left(\frac{x_i}{n}\right)$$
$$= \sum_{i=1}^{n}\frac{1}{n^2}Var(x_i) = \frac{1}{n^2}\sum_{i=1}^{n}\sigma^2 = \frac{1}{n^2}(n\sigma^2)$$
$$= \sigma^2/n$$

10. The expected value of the sample variance is population variance:

$$E(x_i^2) = E((x_i + \mu - \mu)^2) = E((x_i-\mu)^2 - 2\mu(x_i-\mu) + \mu^2)$$
$$= E((x_i-\mu)^2) - 2\mu E(x_i-\mu) + E(\mu^2)$$
$$= \sigma^2 + \mu^2$$
$$E(\bar{x}^2) = E((\bar{x} + \mu - \mu)^2) = E((\bar{x}-\mu)^2 - 2\mu(\bar{x}-\mu) + \mu^2)$$
$$= E((\bar{x}-\mu)^2) - 2\mu E(\bar{x}-\mu) + E(\mu^2)$$
$$= \frac{\sigma^2}{n} + \mu^2$$
$$s^2 = \frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n-1} = \frac{1}{n-1}\left(\sum_{i=1}^{n} x_i^2 - 2\bar{x}\sum_{i=1}^{n} x_i + n\,\bar{x}^2\right)$$
$$= \frac{1}{n-1}\left(\sum_{i=1}^{n} x_i^2 - n\,\bar{x}^2\right)$$

$$E(s^2) = E\left(\frac{1}{n-1}\left(\sum_{i=1}^{n} x_i^2 - n\,\bar{x}^2\right)\right) = \frac{1}{n-1}\left(\sum_{i=1}^{n} E(x_i^2) - nE(\,\bar{x}^2)\right)$$
$$= \frac{1}{n-1}\left(n(\sigma^2+\mu^2) - n\left(\frac{\sigma^2}{n} + \mu^2\right)\right) = \sigma^2$$

Thus, the sample variance is the "unbiased" estimator of the population variance.

Appendix 6. Quartile Method for Determination of Outliers

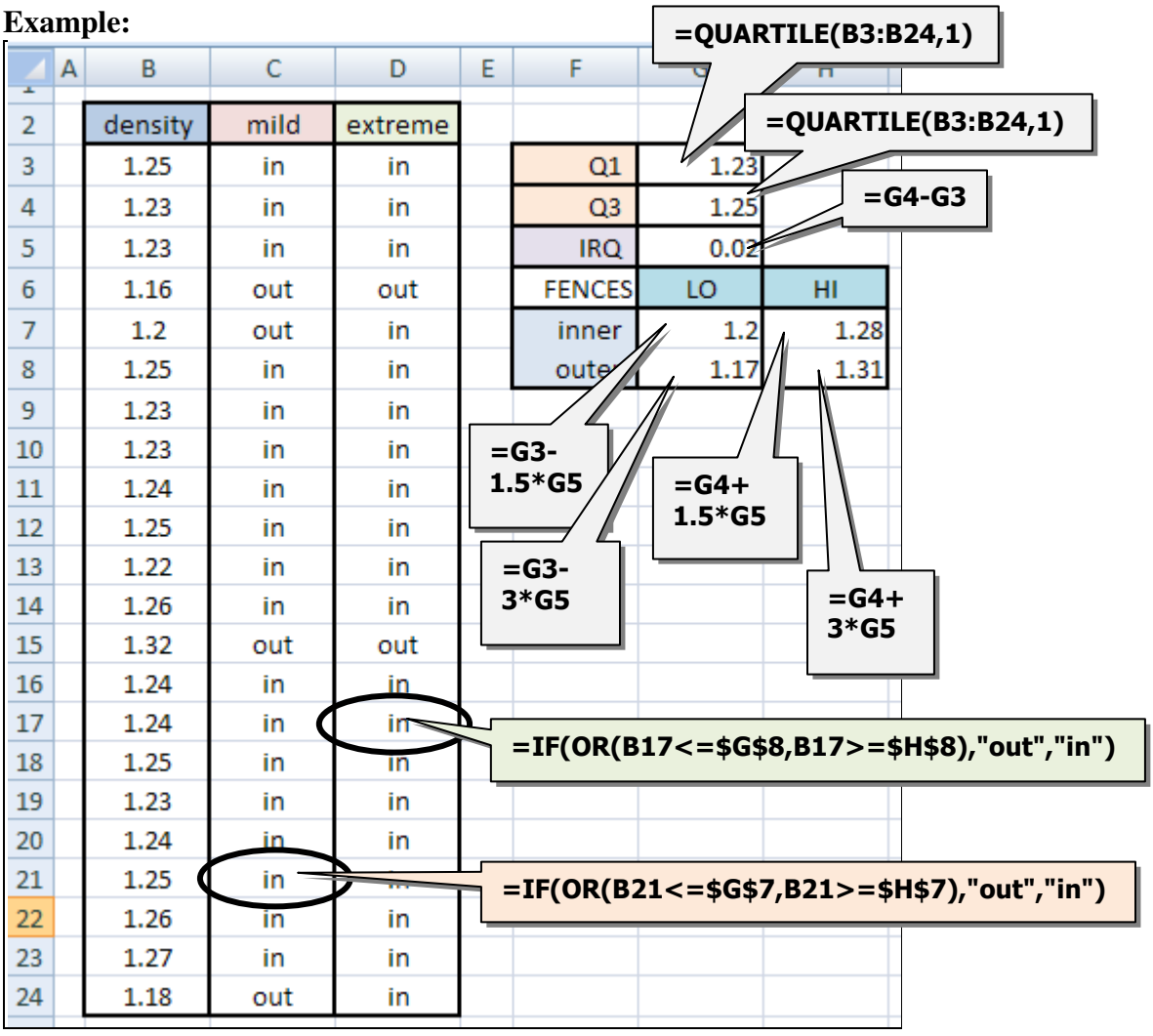
(tbco 10/16/2007)

Method:

Let Q_1 be the first quartile and Q_3be the third quartile.

1. Compute the difference, called the inter-quartile range: $IQR = Q_3 - Q_1$.
2. Then calculate the inner fence:
   Minimum value: $Q_1 - 1.5\,IQR$
   Maximum value: $Q_2 + 1.5\,IQR$
   and the outer fence:
   Minimum value: $Q_1 - 3\,IQR$
   Maximum value: $Q_2 + 3\,IQR$
3. The values outside inner fence are considered "mild outliers", while the values outside the outer fence are considered as "extreme outliers".

Example:



	density	mild	extreme
3	1.25	in	in
4	1.23	in	in
5	1.23	in	in
6	1.16	out	out
7	1.2	out	in
8	1.25	in	in
9	1.23	in	in
10	1.23	in	in
11	1.24	in	in
12	1.25	in	in
13	1.22	in	in
14	1.26	in	in
15	1.32	out	out
16	1.24	in	in
17	1.24	in	in
18	1.25	in	in
19	1.23	in	in
20	1.24	in	in
21	1.25	in	in
22	1.26	in	in
23	1.27	in	in
24	1.18	out	in

Q1	1.23	
Q3	1.25	
IRQ	0.02	
FENCES	LO	HI
inner	1.2	1.28
outer	1.17	1.31

Appendix 7. Histogram Macro in Excel

(tbco 10/17/2007)

Purpose of histograms:

To visualize the count or frequency of data inside chosen intervals known as bins.

Histogram Macro

Note: We have built a macro for the construction of histograms based on intervals. Excel has a built-in function for making histograms but require adjustment of the bar chart to obtain standard histograms.

Downloading

A zipped version of the file **`histogram.bas`** is available for download using the link: www.chem.mtu.edu/~tbco/cm3215/histogram.zip

Activation

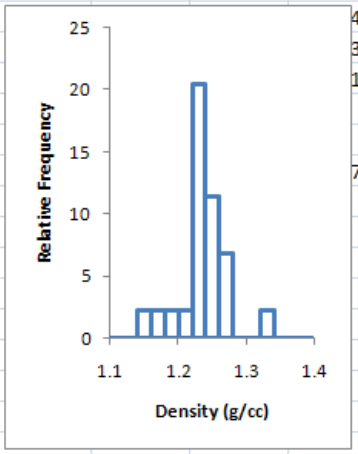
1. With an open excel worksheet, click **[Alt-F11]** to open the VBA (visual basic for applications) editor.
2. Click **[Ctrl-M]** and import the file **`histogram.bas`**.
3. Click **[Alt-F11]** to return to the Excel worksheet.

Using the Macro

1. Make sure data is available in the worksheet.
2. Invoke the histogram macro using **[CTRL-h]**.(Alternatively, you can select **[VIEW]→[Macro]→[View Macros…]** menu item then select **[histogram]** and click **[Run]**.
3. Follow the instructions prompted by the input boxes:
   a. Data range: click-drag to select cells.
   b. Minimum bin value: can be less than the minimum data value
   c. Maximum bin value: can be greater than the maximum data value
   d. Bin interval width: must be a fraction of the range
   e. Select 1 (Frequency count) or 2(relative frequency)
   f. Cell to store results: row location must be greater than row 2.
4. Change the graph if desired, e.g. axis titles and range.

Example:

	A	B	C	D	E	F	G	H	I	J
1	density			Bins	Count Fre	Relative Frequency		xBin	yRelative Frequenc	
2	1.25			1.1	0	0		1.1	0	
3	1.23			1.12	0	0		1.1	0	
4	1.23			1.14	1	2.272727		1.12	0	
5	1.16			1.16	1	2.272727		1.12	0	
6	1.2			1.18	1	2.272727		1.12	0	
7	1.25			1.2	1	2.272727		1.12	0	
8	1.23					455		1.14	0	
9	1.23					364		1.14	0	
10	1.24					182		1.14	0	
11	1.25					0		1.14	2.272727	
12	1.22					0		1.16	2.272727	
13	1.26					727		1.16	0	
14	1.32					0		1.16	0	
15	1.24					0		1.16	2.272727	
16	1.24					0		1.18	2.272727	
17	1.25							1.18	0	
18	1.23							1.18	0	
19	1.24							1.18	2.272727	
20	1.25							1.2	2.272727	
21	1.26							1.2	0	
22	1.27							1.2	0	
23	1.18							1.2	2.272727	
24								1.22	2.272727	



Remarks:

1. For the results shown, we used the following input:
   a. Cell range: $A$2:$A$23
   b. Minimum bin value: 1.1
   c. Maximum bin value: 1.4
   d. Interval width: 0.02
   e. Type: 2
   f. Cell to store results: $D$2
2. The columns labeled **bins**, **frequency count** and **relative frequency** are histogram analysis results.
3. The columns labeled x**bins** and **yrelative frequency** are just used for plotting the histogram.
4. The axis range and titles of the plot were then modified manually.

Appendix 8. Normal Quantile Plots

(tbco 10/18/2007)

Purpose:

To check whether a data is normally distributed.
(Remark: histograms of small data sets can be very sensitive to the choice of bin width, but cumulative frequency information is more robust.)

Method:

1. Arrange the data in ascending order: $d_1 \le d_2 \le \cdots \le d_n$.
2. Calculate the corresponding quantile: $q_i = {(i - 0.5)}/{n}$.
3. Determine the normal score z_i that would yield this cumulative frequency.
$$z_i = \text{NORMSINV}(q_i)$$
i.e., the inverse standard normal cumulative distribution function of q_i.
4. Plot data versus the normal score, e.g. z_i vs d_i
5. If the plot lie close to a line that passes through z=0, then the data is considered close to normally distributed.

Example:

Consider the following data:

3.80	3.70	3.57	3.41	3.58	3.33
3.49	3.56	3.50	3.50	3.40	3.37
3.75	3.51	3.52	3.25	3.60	3.60
3.62	3.51	3.38	3.40	3.42	3.34
3.47	3.61	3.55	3.54	3.20	3.46

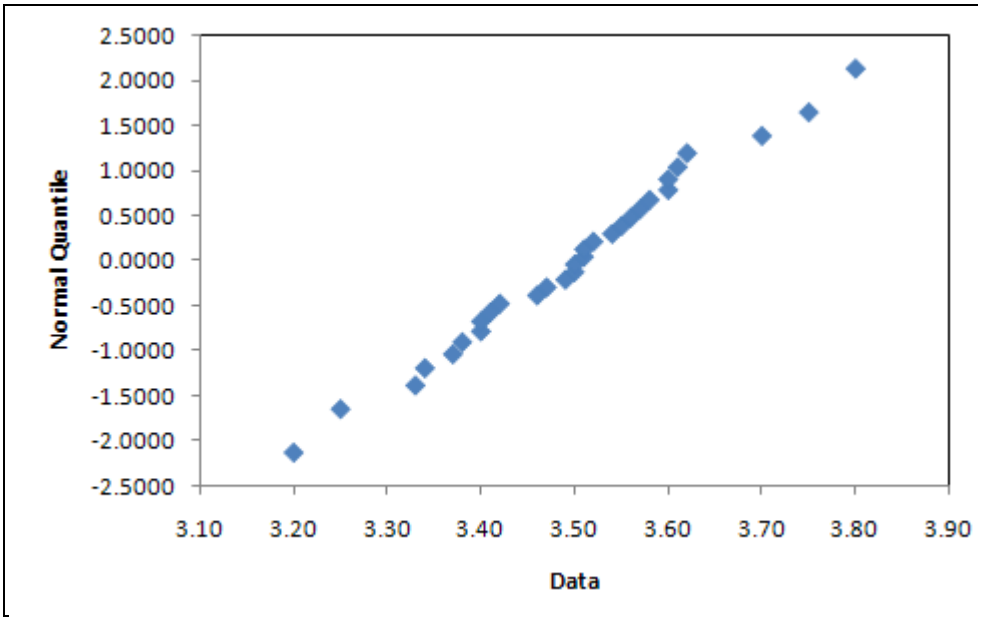
Then a spreadsheet could be constructed as follows:

	A	B	C	D	E	F
1	Data		Sorted	Rank	Quantile	N-Score
2	3.80		3.20	1	0.0167	-2.1280
3	3.49		3.25	2	0.0500	-1.6449
4	3.75		3.33	3	0.0833	-1.3830
5	3.62		3.34	4	0.1167	-1.1918
6	3.47				0.1500	.0364
27	3.58		3.61	26	0.8500	.0364
28	3.40		3.62	27	0.883	1.1918
29	3.60		3.70			30
30	3.42		3.75	29	0.9500	1.6449
31	3.20		3.80	30	0.9833	2.1280

=(D2-0.5)/30

=NORMSINV(E5)

and then plot column F vs column c:



The data fall pretty much on a line, thus we can conclude that the data is normally distributed.

Appendix 9. Plotting Normal Distribution Curve Using Excel

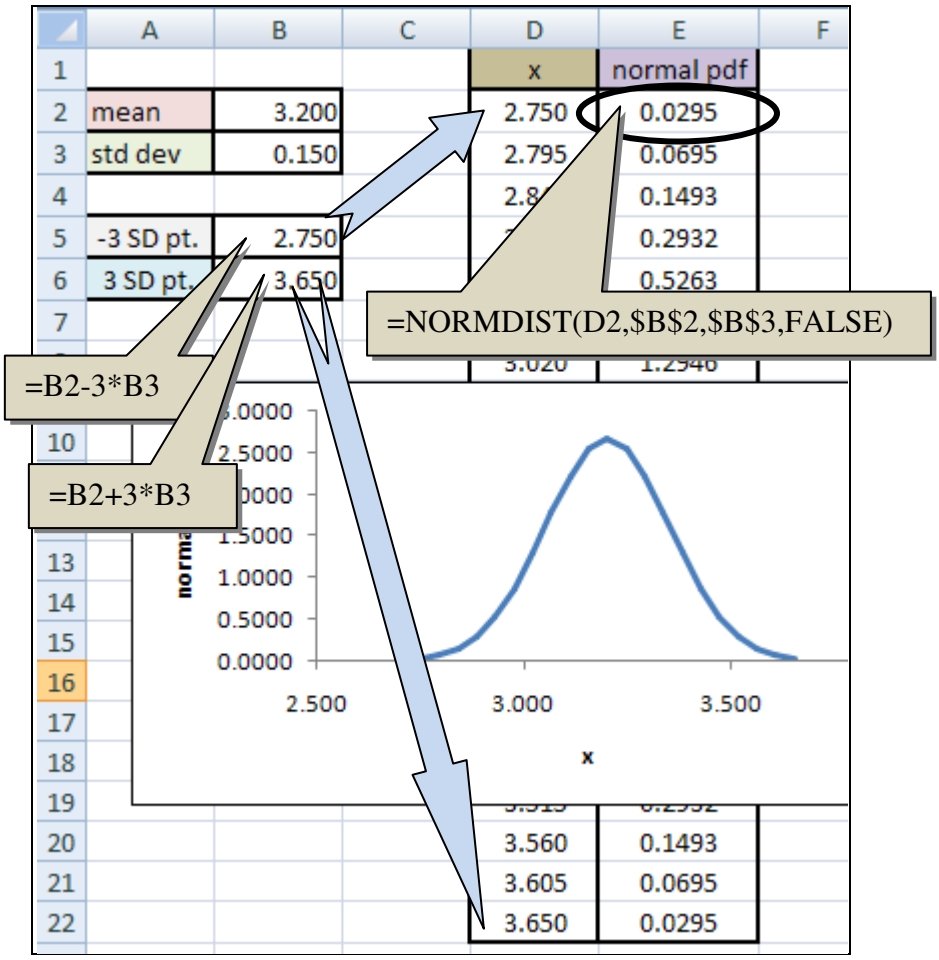
(tbco 10/19/2007)

Purpose:

To plot a normal distribution curve (probability distribution function) based on the given values of mean and standard deviation.

Procedure:

1. First set up cells containing mean and standard deviation
2. Calculate a range of values of the random variable. For example, you can first evaluate the values ranging from three sigma below the mean and three sigma values above the mean.
3. Use the Excel function NORMDIST(x, mean, sd, FALSE) to calculate the pdf values. (Note: the FALSE value is to set the mode to pdf, otherwise it yields cumulative frequencies)
4. Plot the pdf vs. the random variable.



Appendix 10. Confidence Intervals Using Excel

(tbc 10/20/2007)

Purpose:

Using built-in functions, we can obtain the confidence interval of the population mean based on small samples using the t-distribution.

Method:

1. Calculate the mean, $\bar{x}$, and standard deviation, s, using the built-in functions **AVERAGE** and **STDEV**.
2. Calculate the standard error (SE),

$$SE = \frac{s}{\sqrt{n}}$$

3. Determine the scaling factor of the standard error that would yield the desired confidence level.

   For example, for a 95% confidence interval:

   a. Small sample size ($n < 20$):

   $$\text{scaling factor} = t_{95} = \text{TINV}(0.05, n-1)$$

   b. Large sample size:

   $$\text{scaling factor} = n_{95} = \text{NORMSINV}(0.025)$$

4. Evaluate the lower limit and upper limit of the confidence interval,

$$\text{Lower limit} = \bar{x} - (\text{scaling factor})(SE)$$
$$\text{Upper limit} = \bar{x} + (\text{scaling factor})(SE)$$

5. If desired, plot the mean together with the confidence limits using error bars.

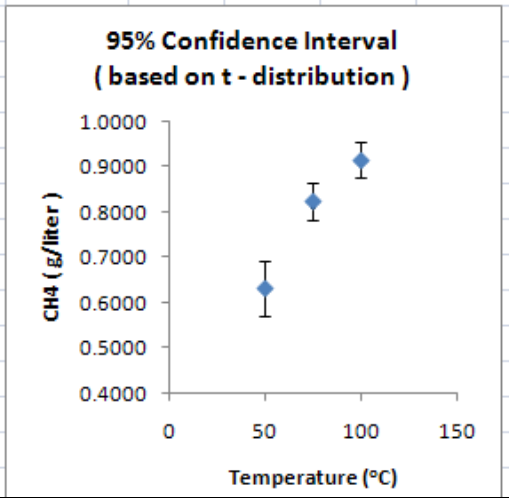
Caution:

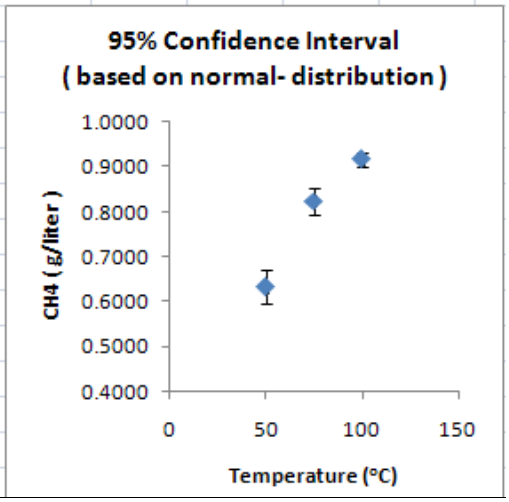
The usual convention for error bars is to plot: $\bar{x} \pm SE$. This means a 68.3% confidence interval based on normal distributions. For small samples, this leads to even lower confidence levels since *t*-distributions are needed. However, due to the difficulty of obtaining inverse *t*-distributions in the past (requiring interpolations from table of *t*-distribution data), the practice of just using the standard error SE is common and accepted. Below, we show an example of plotting the 95% confidence interval in terms of error bars. When doing so, please indicate that these are "estimates with 95% confidence intervals". Otherwise, most scientists and statisticians will probably misinterpret your plots.

Example:

Temp (°C)	Concentration of CH4 (g/liter)					n	mean	st dev	S Error
50	0.65	0.63	0.58	0.67		4	0.6325	0.0386	0.01931
75	0.82	0.81	0.82	0.88	0.79	5	0.8240	0.0336	0.01503
100	0.93	0.9	0.92			3	0.9167	0.0153	0.00882

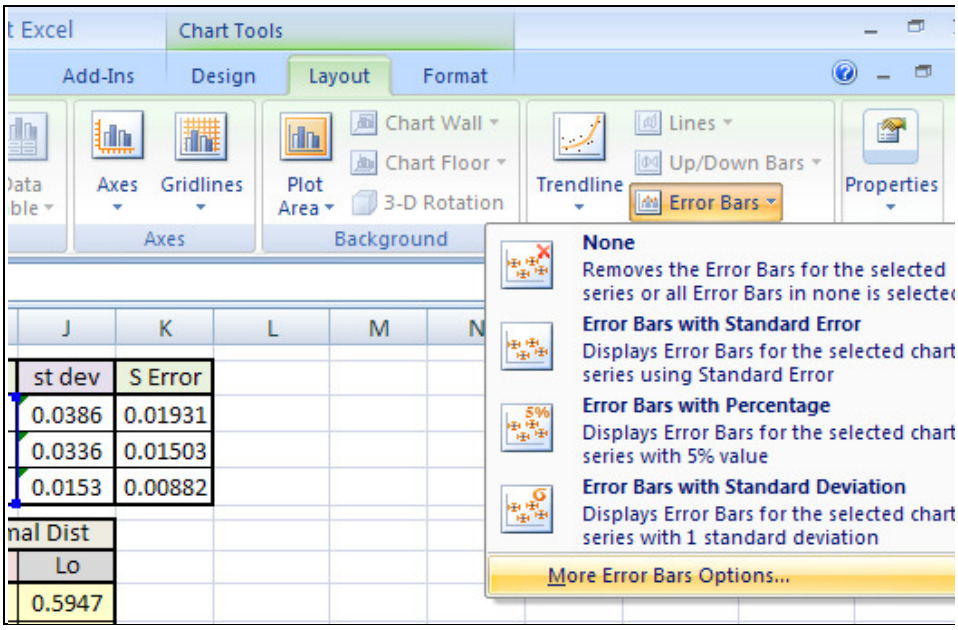
		Based on t Dist				Based on Normal Dist		
Temp (°C)	t 95	width/2	Hi	Lo	n 95	width/2	Hi	Lo
50	3.1824	0.06146	0.6940	0.5710	1.9600	0.03785	0.6703	0.5947
75	2.7764	0.04174	0.8657	0.7823	1.9600	0.02946	0.8535	0.7945
100	4.3027	0.03795	0.9546	0.8787	1.9600	0.01729	0.9340	0.8994



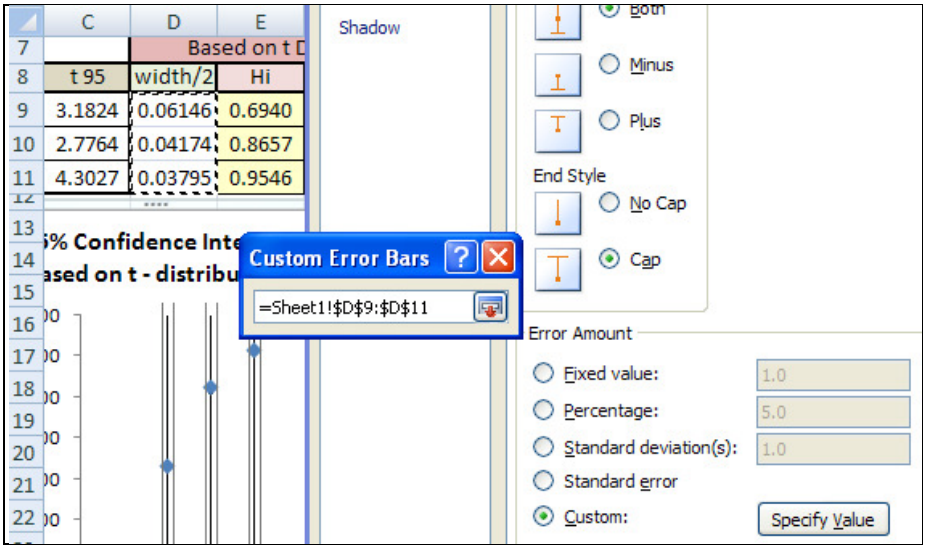


How error bars were included in the plots above:

1. Do an x-y (scatter) plot mean vs. temperature.
2. In the toolbar, select **[Chart Tools]→[Layout]→[Error Bars]→[More Error Bars Options…]**



3. In the pop-up window, choose the **[custom]** selection, then click **[Specify Value]** button. For the *t*-distribution case, we selected the range **[$D$9:$D$11]** for both positive and negative errors values. The error bars should now appear.



References

1. L. Gonick and W. Smith. “Cartoon Guide to Statistics”. Harper Collins Publishers Inc. New York, NY. 1993.
2. C. Mack. “Essentials of Statistics for Scientist and Technologists”. Plenum Press. New York, NY. 1966.
3. G. Geoffrey Vining. “Statistical Methods for Engineers”. Brooks Cole Publishing Co. Pacific Grove, CA. 1998.
4. S. Meyer. “Data Analysis for Scientists and Engineers”. J. Wiley and Sons, Inc. New York, NY. 1975.