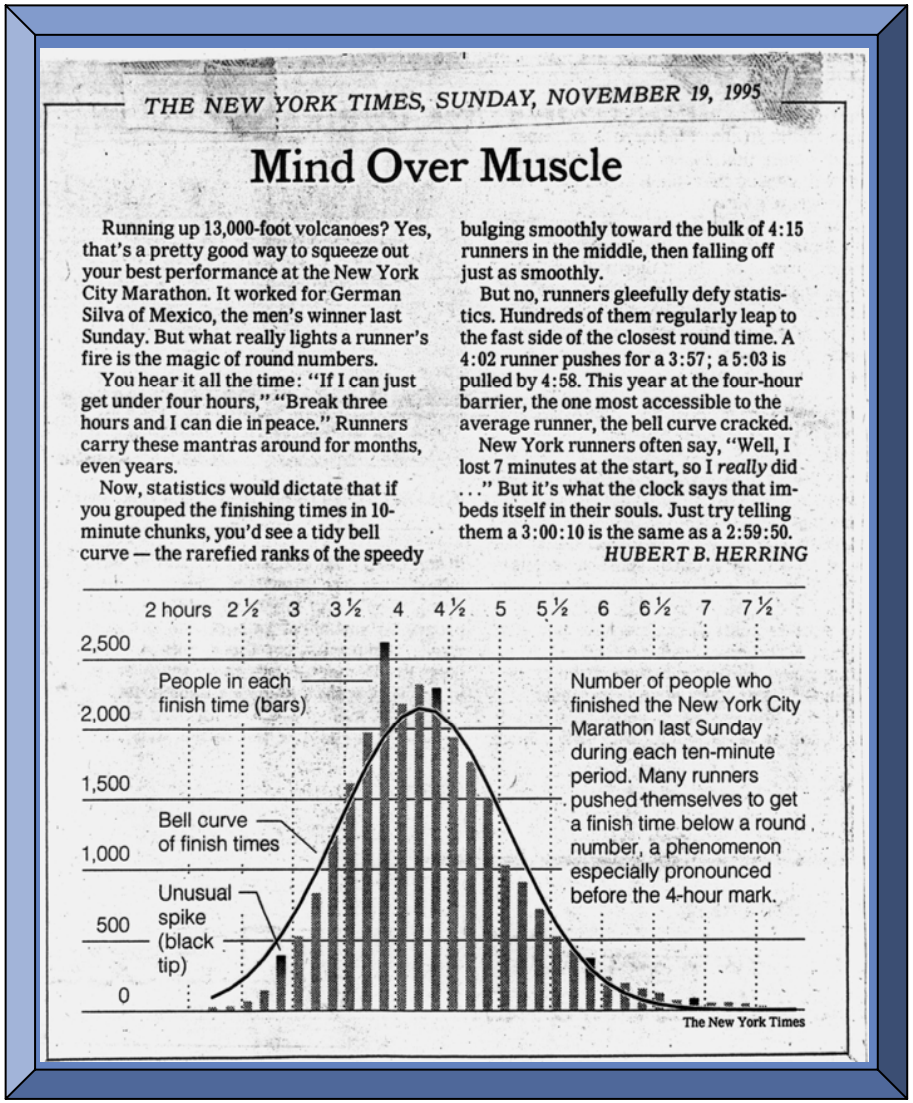
Statistics



THE NEW YORK TIMES, SUNDAY, NOVEMBER 19, 1995

Mind Over Muscle

Running up 13,000-foot volcanoes? Yes, that's a pretty good way to squeeze out your best performance at the New York City Marathon. It worked for German Silva of Mexico, the men's winner last Sunday. But what really lights a runner's fire is the magic of round numbers.

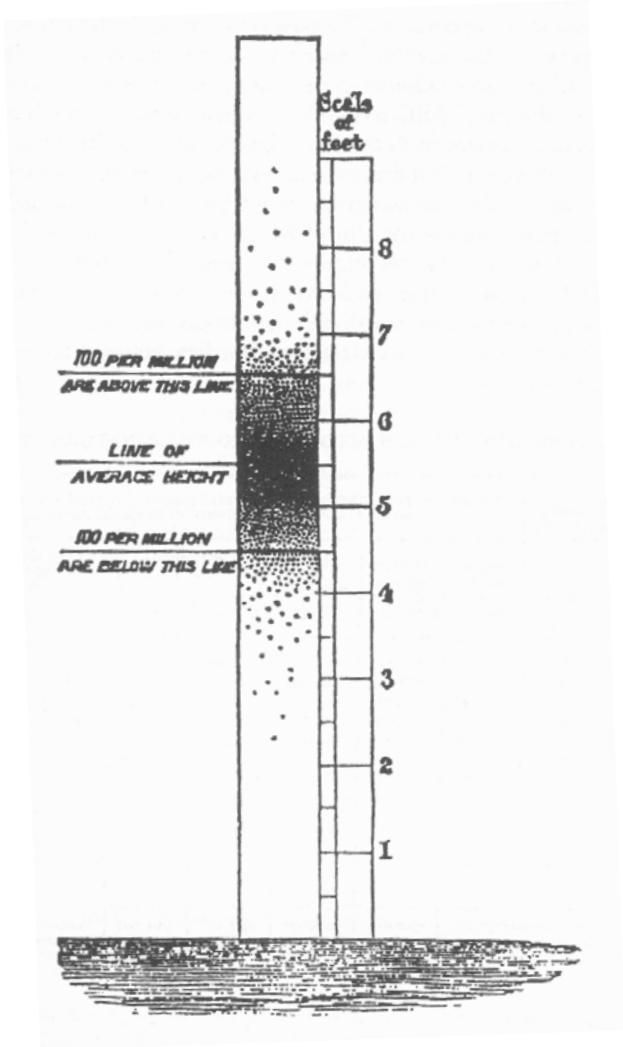
You hear it all the time: "If I can just get under four hours," "Break three hours and I can die in peace." Runners carry these mantras around for months, even years.

Now, statistics would dictate that if you grouped the finishing times in 10-minute chunks, you'd see a tidy bell curve — the rarefied ranks of the speedy bulging smoothly toward the bulk of 4:15 runners in the middle, then falling off just as smoothly.

But no, runners gleefully defy statistics. Hundreds of them regularly leap to the fast side of the closest round time. A 4:02 runner pushes for a 3:57; a 5:03 is pulled by 4:58. This year at the four-hour barrier, the one most accessible to the average runner, the bell curve cracked.

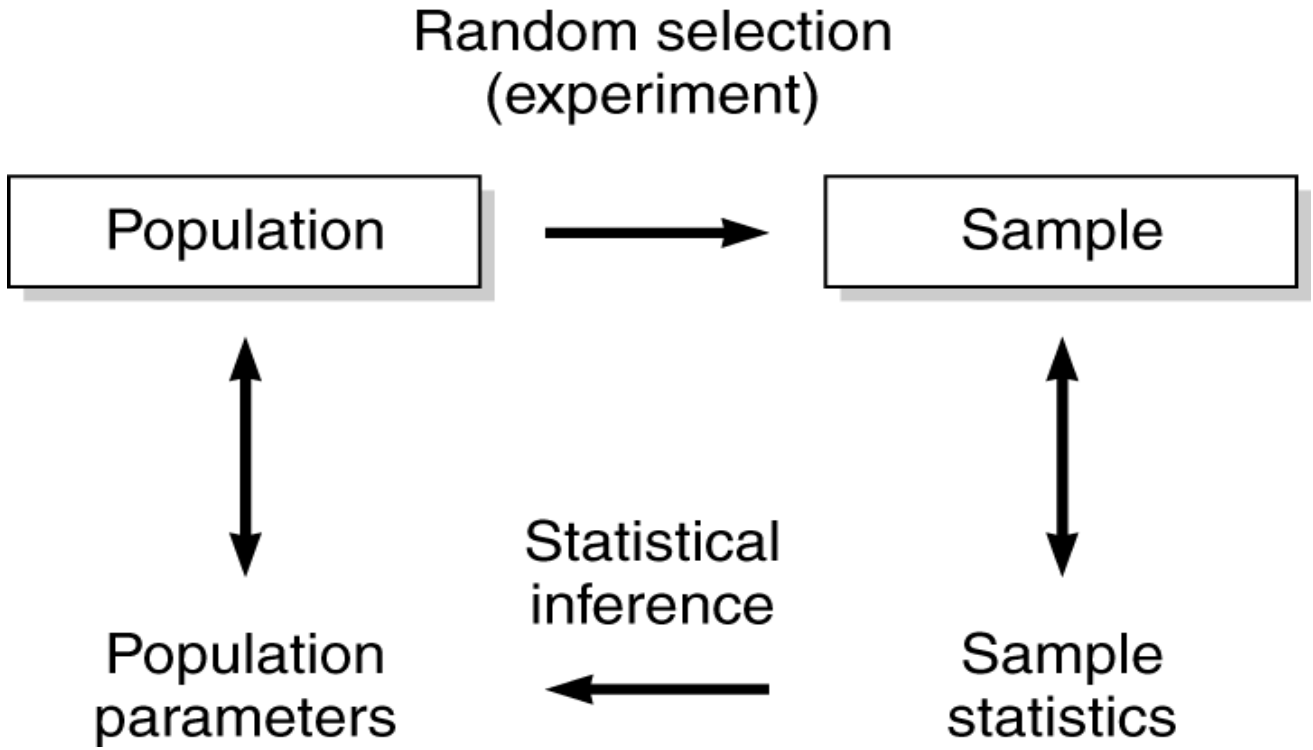
New York runners often say, "Well, I lost 7 minutes at the start, so I *really* did . . ." But it's what the clock says that imbeds itself in their souls. Just try telling them a 3:00:10 is the same as a 2:59:50.

HUBERT B. HERRING



Galton's Heights of Hypothetical Men (1869)

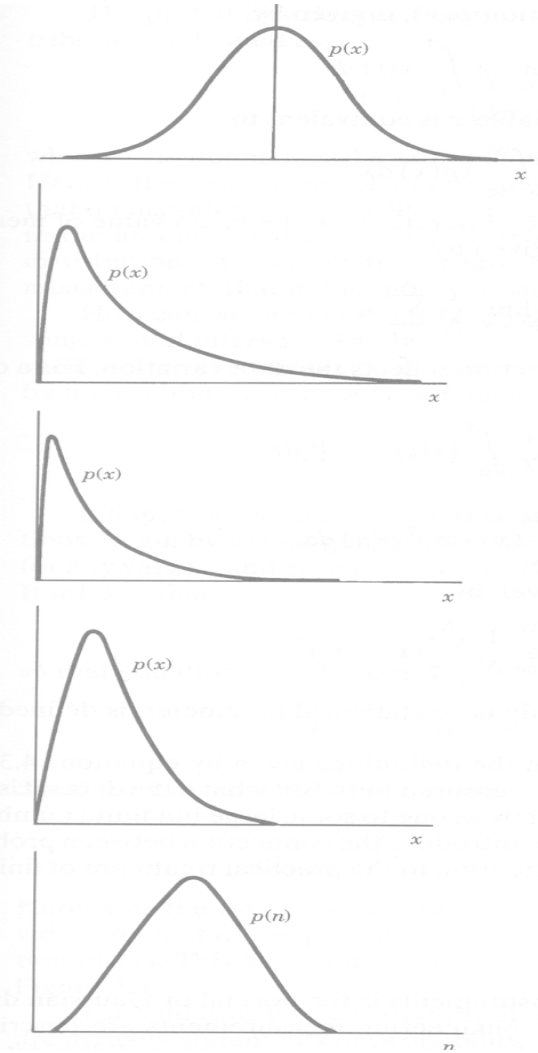
Statistical Inference



$$x' \Leftarrow \bar{x} - t_{\nu,P}\frac{S_x}{\sqrt{N}} \le x' \le \bar{x} + t_{\nu,P}\frac{S_x}{\sqrt{N}} \Leftarrow \bar{x}$$

$$\sigma^2 \Leftarrow \nu S_x^2 / \chi^2_{\alpha/2} \le \sigma^2 \le \nu S_x^2 / \chi^2_{1-\alpha/2} \Leftarrow S_x^2$$

Some Probability Density Functions



Normal: outcome influenced by a very large number of very small, ‘50-50 chance’ effects (Ex: human heights)

Lognormal: outcome influenced by a very large number of very small, ‘constrained’ effects (Ex: rain drops)

Poisson: outcome influenced by a rarely occurring events in a very large population (Ex: micrometeoroid diameters in LEO)

Weibull: outcome influenced by a ‘failure’ event in a very large population (Ex: component life time)

Binomial: outcome influenced by a finite number of ’50-50 chance’ effects (Ex: coin toss)

The Normal Distribution

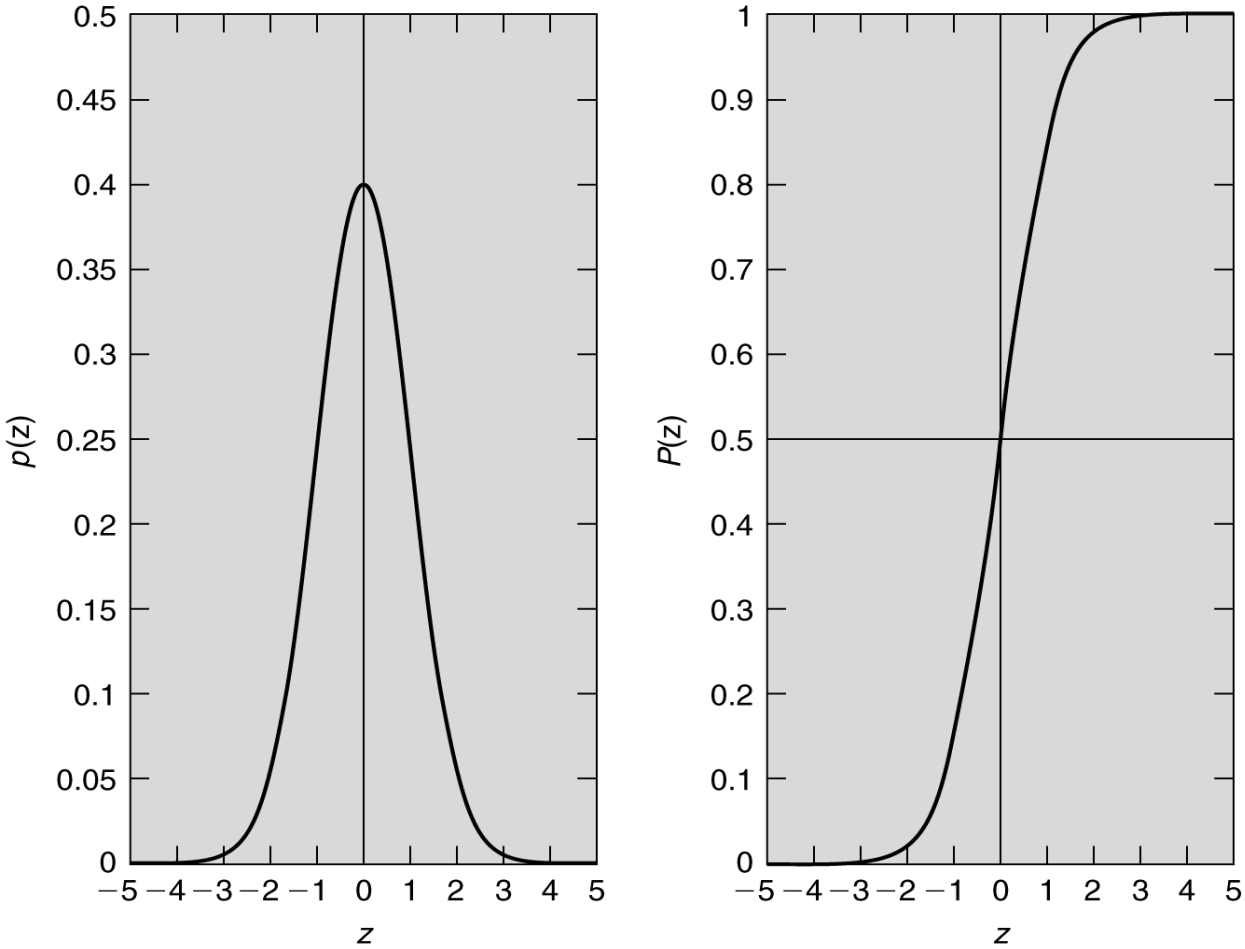
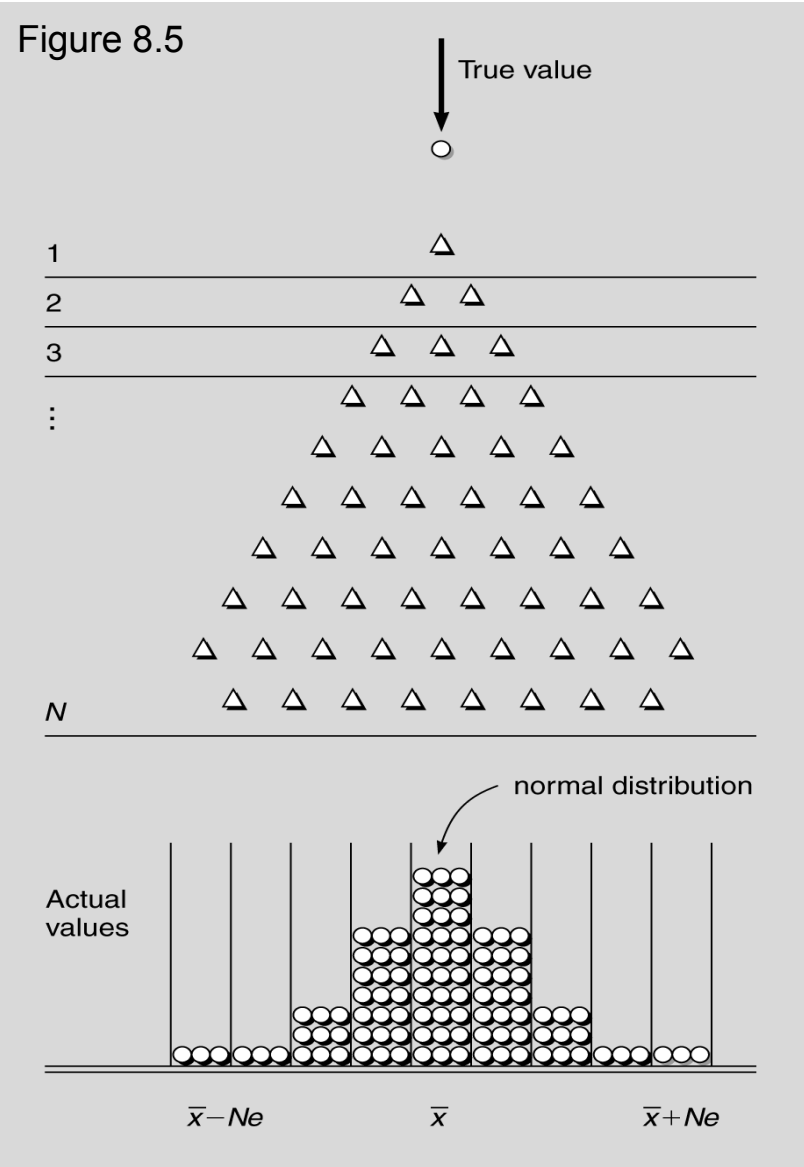


Figure 8.4

Concept of the Normal Distribution



Population, with true mean and true variance x‘ and σ

Experiment with many, small, uncontrolled extraneous variables

Normalized Variables

$$p(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{1}{2\sigma^2}(x - x')^2 \right]$$

$\beta = (x\text{-}x') / \sigma$ standardized normal variable

$z_1 = (x_1\text{-}x') / \sigma$ normalized *z*-variable (z_1 is a *specific* value of β); subscript 1 usually dropped

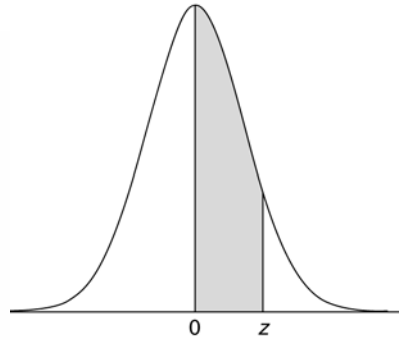
$$\Pr\left[-z_1 \le \beta \le +z_1 \right] = \frac{1}{\sqrt{2\pi}} \int_{-z_1}^{+z_1} \exp\left[-\frac{\beta^2}{2} \right] d\beta = 2p(z_1)$$

where $p(z_1)$ is the **normal error function**.

Normal Error Function Table

ONE-SIDED NORMAL ERROR FUNCTION VALUES

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	.0000	.0040	.0080	.0120	.0160	.0199	.0239	.0279	.0319	.0359
0.1	.0398	.0438	.0478	.0517	.0557	.0596	.0636	.0675	.0714	.0753
0.2	.0793	.0832	.0871	.0910	.0948	.0987	.1026	.1064	.1103	.1141
0.3	.1179	.1217	.1255	.1293	.1331	.1368	.1406	.1443	.1480	.1517
0.4	.1554	.1591	.1628	.1664	.1700	.1736	.1772	.1808	.1844	.1879
0.5	.1915	.1950	.1985	.2019	.2054	.2088	.2123	.2157	.2190	.2224
0.6	.2257	.2291	.2324	.2357	.2389	.2422	.2454	.2486	.2517	.2549
0.7	.2580	.2611	.2642	.2673	.2704	.2734	.2764	.2794	.2823	.2852
0.8	.2881	.2910	.2939	.2967	.2995	.3023	.3051	.3078	.3106	.3133
0.9	.3159	.3186	.3212	.3238	.3264	.3289	.3315	.3340	.3365	.3389
1.0	.3413	.3438	.3461	.3485	.3508	.3531	.3554	.3577	.3599	.3621
1.1	.3643	.3665	.3686	.3708	.3729	.3749	.3770	.3790	.3810	.3830
1.2	.3849	.3869	.3888	.3907	.3925	.3944	.3962	.3980	.3997	.4015
1.3	.4032	.4049	.4066	.4082	.4099	.4115	.4131	.4147	.4162	.4177
1.4	.4192	.4207	.4222	.4236	.4251	.4265	.4279	.4292	.4306	.4319
1.5	.4332	.4345	.4357	.4370	.4382	.4394	.4406	.4418	.4429	.4441
1.6	.4452	.4463	.4474	.4484	.4495	.4505	.4515	.4525	.4535	.4545
1.7	.4554	.4564	.4573	.4582	.4591	.4599	.4608	.4616	.4625	.4633
1.8	.4641	.4649	.4656	.4664	.4671	.4678	.4686	.4693	.4699	.4706
1.9	.4713	.4719	.4726	.4732	.4738	.4744	.4750	.4758	.4761	.4767
2.0	.4772	.4778	.4783	.4788	.4793	.4799	.4803	.4808	.4812	.4817
2.1	.4821	.4826	.4830	.4834	.4838	.4842	.4846	.4850	.4854	.4857
2.2	.4861	.4864	.4868	.4871	.4875	.4878	.4881	.4884	.4887	.4890
2.3	.4893	.4896	.4898	.4901	.4904	.4906	.4909	.4911	.4913	.4916
2.4	.4918	.4920	.4922	.4925	.4927	.4929	.4931	.4932	.4934	.4936
2.5	.4938	.4940	.4941	.4943	.4945	.4946	.4948	.4949	.4951	.4952
2.6	.4953	.4955	.4956	.4957	.4959	.4960	.4961	.4962	.4963	.4964
2.7	.4965	.4966	.4967	.4968	.4969	.4970	.4971	.4972	.4973	.4974
2.8	.4974	.4975	.4976	.4977	.4977	.4978	.4979	.4979	.4980	.4981
2.9	.4981	.4982	.4982	.4983	.4984	.4984	.4985	.4985	.4986	.4986
3.0	.4987	.4987	.4987	.4988	.4988	.4988	.4989	.4989	.4989	.4990
3.5	.4998	.4998	.4998	.4998	.4998	.4998	.4998	.4998	.4998	.4998
4.0	.5000	.5000	.5000	.5000	.5000	.5000	.5000	.5000	.5000	.5000



Pr[0≤z≤1] =

Pr[-1≤z≤1] =

Pr[-2≤z≤2] =

Pr[-3≤z≤3] =

Pr[-0.44≤z≤4.06] =

Table 8.2

In-Class Problem

- What is the probability that a student will score between 75 and 90 on an exam, assuming that the scores are distributed normally with a mean of 60 and a standard deviation of 15 ?

Degrees of Freedom

- The number of **degrees of freedom**, ν, equal the number of data points, *N*, minus the number of independent restrictions (constraints), *c*, used for the required calculations.

$$\nu =$$

- When computing the sample mean, ν =

- When computing the sample standard deviation, ν =

- When constructing a histogram, ν = , for *K* bins .

- When performing a regression analysis, ν = ,
  where m is the order of the fit (linear, m = 1).

Statistics Using MATLAB®

- MATLAB®'s statistics toolbox contains distributions such as the normal (*norm*), Student's t (*t*), and χ^2 (*chi2*).

- Some useful command prefixes are the probability density function (*pdf*), probability distribution function (*cdf*), and the inverse of the distributions (*inv*).

- The command *normpdf(x,x',σ)* gives the value of p(x).

  normpdf(0,0,1) = 0.3989

- The command *normcdf(x*,x',σ)* gives $\mathrm{P(x^*)} = \int_{-\infty}^{x^*} \mathrm{p(x)dx}$

  normcdf(1,0,1) = 0.8413 (= 0.3413 + 0.5000)

  normal error function, $p(z_1)$ = normcdf(z_1,0,1) – 0.5

- The command *norminv(P,x',σ)* gives x* from the cdf

  norminv(0.8413,0,1) = 1

Statistics Using MATLAB® (cont'd)

- To obtain the value of z for a given %P, x' and σ:

$$z = \frac{\text{norminv}\left(\frac{1+P}{2}, x', \sigma\right) - \text{norminv}\left(\frac{1-P}{2}, x', \sigma\right)}{2}$$

- z(95,0,1) = 1.96 and z(95.45,0,1) = 2.00