Statistics

Statistics Defined

- Statistics is the study of the analysis and interpretation of empirical data

Statistical Sampling

- In cases in which the amount of data available for analysis is very large a **sample** is taken from the total **population** of data which is usually much larger than the sample and is often infinite or so large as to be practically infinite
- If the sample is properly taken its characteristics are representative of the characteristics of the population

Sample Mean and Variance

- A **descriptor** is a number that generally characterizes a random variable
- The **mean** is the most important descriptor
- The population mean is the mean of all the available data
- The sample mean is the mean of the data in a sample and is an **estimator** of the population mean

Sample Mean and Variance

The population mean is the limit of the sample mean as the sample size approaches infinity

$$\mu_X = \lim_{N\to\infty}\frac{1}{N}\sum_{i=1}^{N} x_i$$

X

⋮
-0.4650
0.3710
0.7283
2.1122
-1.3573
-1.0226
1.0378
-0.3898
-1.3813
0.3155
1.5532
0.7079
1.9574
0.5045
1.8645
-0.3398
-1.1398
-0.2111
1.1902
-1.1162
0.6353
-0.6014
0.5512
-1.0998
0.0860
-2.0046
-0.4931
0.4620
-0.3210
1.2366
⋮

Sample Mean as a Constant

$\bar{x} = 0.4393$

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

X

⋮
-0.4650
0.3710
0.7283
2.1122
-1.3573
-1.0226
1.0378
-0.3898
-1.3813
0.3155
1.5532
0.7079
1.9574
0.5045
1.8645
-0.3398
-1.1398
-0.2111
1.1902
-1.1162
0.6353
-0.6014
0.5512
-1.0998
0.0860
-2.0046
-0.4931
0.4620
-0.3210
1.2366
⋮

$\bar{X}$

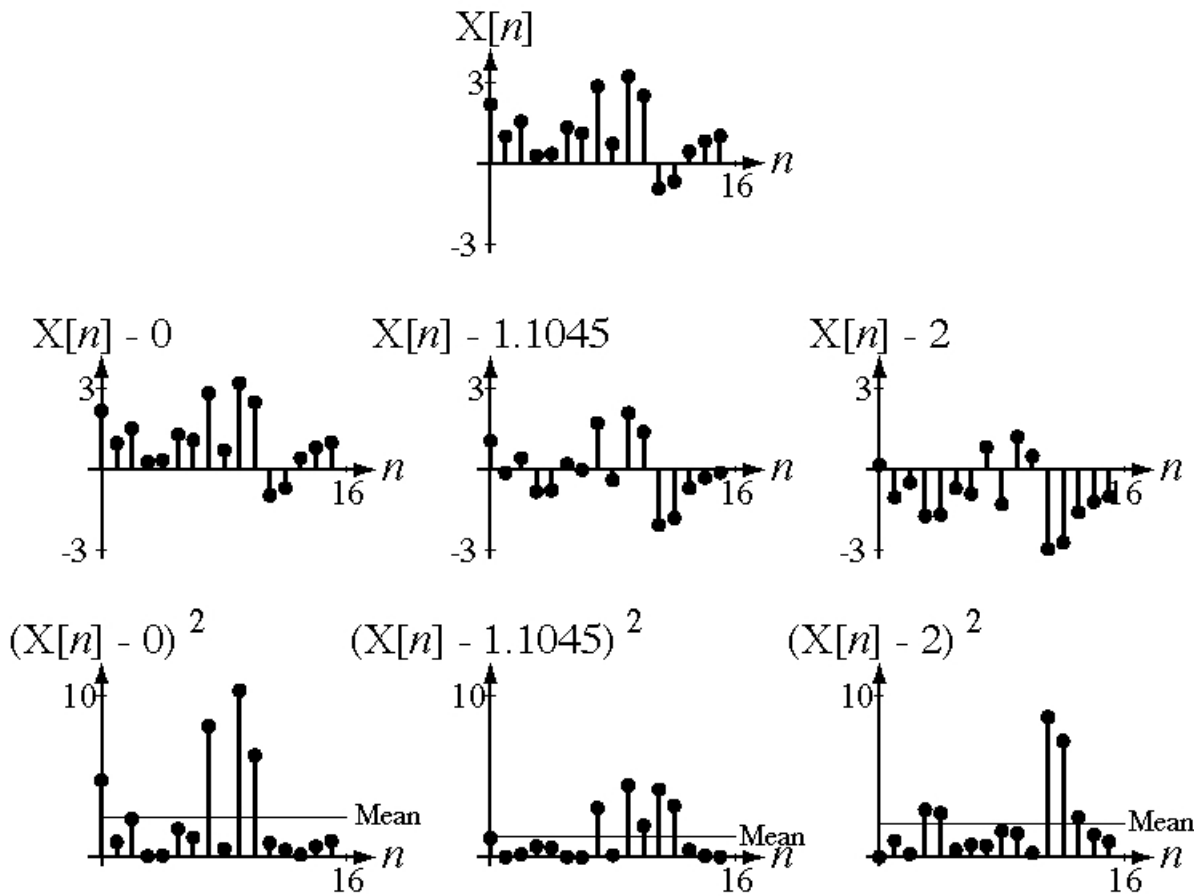
⋮
0.0611
0.6095
-0.07345
-0.17235
⋮

Sample Mean as a Random Variable

$$\bar{X} = \frac{1}{N}\sum_{i=1}^{N} X_i$$

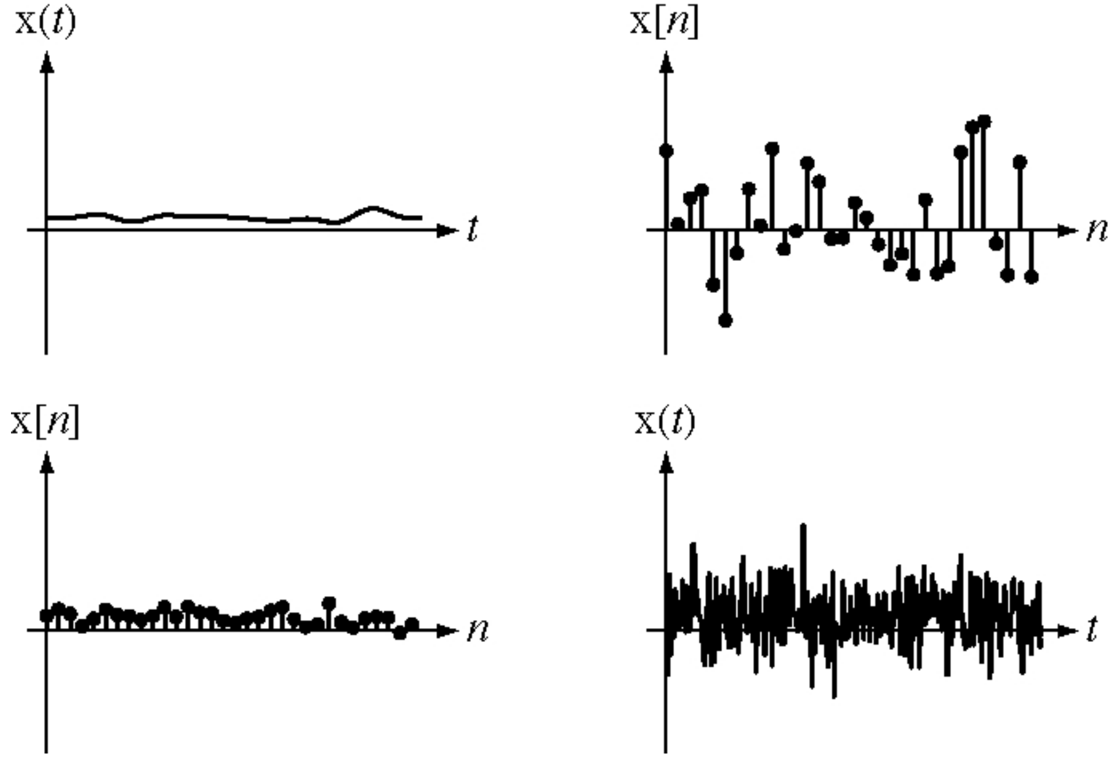
Sample Mean and Variance

The sample mean is the number that minimizes the sum of the squared differences between it and the sample values



Sample Mean and Variance

In this illustration the four signals all have the same mean value.



After the mean, the next most important descriptor is the **standard deviation**. The standard deviation indicates generally how much a signal deviates from its mean value.

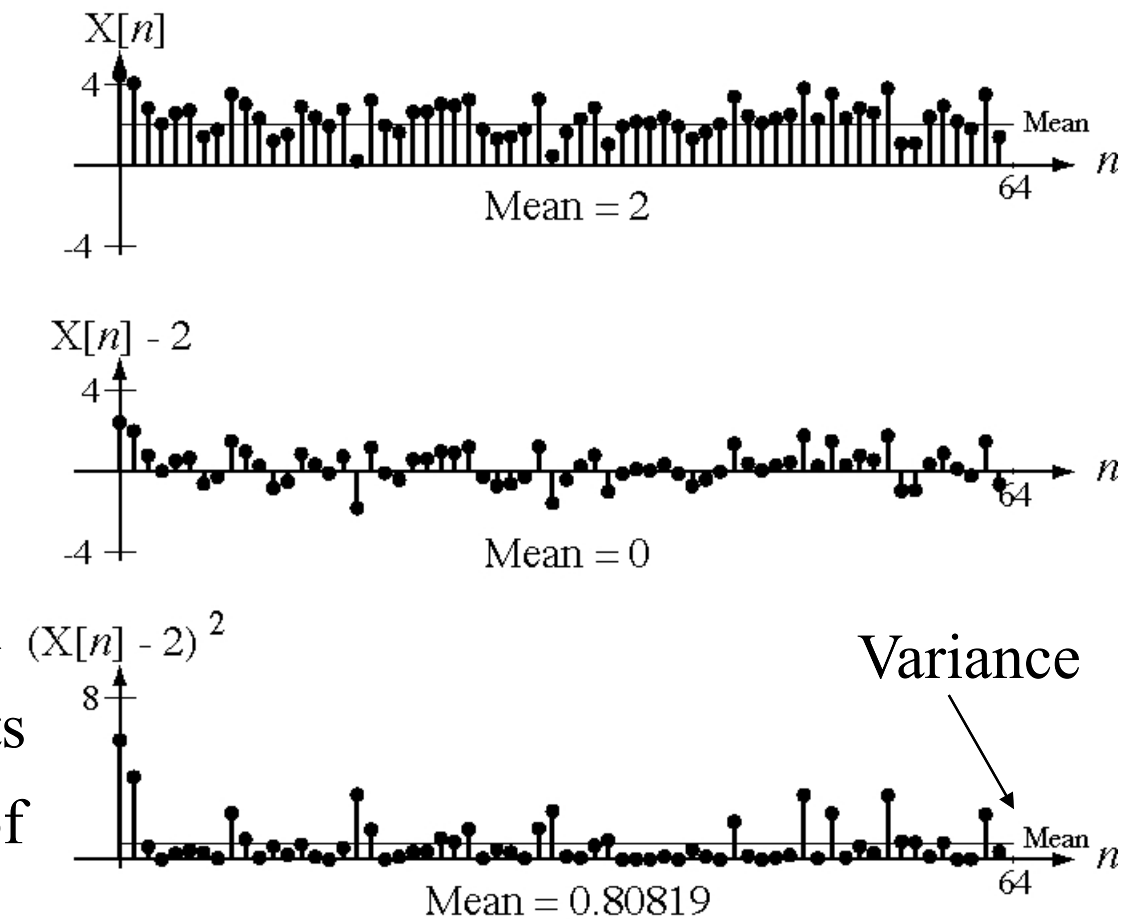
Sample Mean and Variance

The standard deviation is defined by

$$\sigma_X = \sqrt{\mathrm{E}\left(\left|X - \mathrm{E}(X)\right|^2\right)}$$

$$= \sqrt{\lim_{N \to \infty} \frac{1}{N} \sum_{i=1}^{N} \left|x_i - \mu_X\right|^2}$$

It is the square root of the expected value of the squared deviation of the signal from its expected value. The square of the standard deviation is the **variance**.



Sample Mean and Variance

Variance is defined by

$$\sigma_X^2 = \mathrm{E}\left(\left| X - \mathrm{E}(X) \right|^2 \right) = \lim_{N \to \infty} \frac{1}{N} \sum_{n=1}^{N} \left| x_n - \mu_X \right|^2$$

Covariance is a generalization of variance to apply to two different random variables and is defined by

$$\sigma_{XY} = \mathrm{E}\left(\left[X - \mathrm{E}(X) \right] \left[Y - \mathrm{E}(Y) \right]^* \right)$$

which can be expressed as

$$\sigma_{XY} = \mathrm{E}\left(XY^* \right) - \mathrm{E}(X)\mathrm{E}\left(Y^* \right)$$

If X and Y are uncorrelated,

$$\mathrm{E}\left(XY^* \right) = \mathrm{E}(X)\mathrm{E}\left(Y^* \right) \text{ and } \sigma_{XY} = 0$$

Sample Mean and Variance

If variance or mean-squared value or covariance are to be estimated from a finite set of data for two random variables X and Y, they can also be formulated as vector operations. Let the vector of X values be $\mathbf{x}$ and the vector of Y values be $\mathbf{y}$

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_N \end{bmatrix} \ , \ \mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix}$$

Then the mean-squared value of X can be estimated by

$$\mathrm{E}\left(X^2\right) \cong \frac{\mathbf{x}^H \mathbf{x}}{N}$$

where the notation $\mathbf{x}^H$ means the complex conjugate of the transpose of $\mathbf{x}$.

Sample Mean and Variance

The variance of X can be estimated by

$$\sigma_X^2 \cong \frac{\left[\mathbf{x}-\mu_X\right]^H \left[\mathbf{x}-\mu_X\right]}{N}$$

The covariance of X and Y can be estimated by

$$\sigma_{XY} \cong \frac{\left[\mathbf{x}-\mu_X\right]^H \left[\mathbf{y}-\mu_Y\right]}{N}$$

Sample Mean and Variance

The second sense of "sample mean" $\bar{X}$ is itself a random variable and, as such, has a mean and a standard deviation. Its expected value is

$$\mathrm{E}\left(\bar{X}\right) = \mathrm{E}\left(\frac{1}{N}\sum_{n=1}^{N} X_n\right) = \frac{1}{N}\mathrm{E}\left(\sum_{n=1}^{N} X_n\right)$$

$$= \frac{1}{N}\sum_{n=1}^{N}\mathrm{E}\left(X_n\right) = \frac{1}{N} N\,\mathrm{E}\left(X\right) = \mathrm{E}\left(X\right)$$

Since the expected value of the sample mean of X is the same as the expected value of X itself, it is an **unbiased estimator** of the expected value of X.

Sample Mean and Variance

The variance of the sample mean is

$$\sigma_{\bar{X}}^2 = \mathrm{E}\left(\left[\bar{X} - \mathrm{E}\left(\bar{X}\right)\right]\left[\bar{X} - \mathrm{E}\left(\bar{X}\right)\right]^*\right)$$

$$= \mathrm{E}\left(\bar{X}\bar{X}^* - \mathrm{E}\left(\bar{X}\right)\bar{X}^* - \bar{X}\,\mathrm{E}\left(\bar{X}\right)^* + \mathrm{E}\left(\bar{X}\right)\mathrm{E}\left(\bar{X}\right)^*\right)$$

$$= \mathrm{E}\left(\left(\frac{1}{N}\sum_{n=1}^{N} X_n\right)\left(\frac{1}{N}\sum_{m=1}^{N} X_m^*\right)\right) - \left|\mathrm{E}\left(X\right)\right|^2$$

$$= \mathrm{E}\left(\frac{1}{N^2}\sum_{n=1}^{N}\sum_{m=1}^{N} X_n X_m^*\right) - \left|\mathrm{E}\left(X\right)\right|^2$$

Sample Mean and Variance

If X_n and X_m are independently chosen at random from the population they are statistically independent (when $n \neq m$) and

$$\sigma_{\bar{X}}^2 = \frac{1}{N^2} \sum_{n=1}^{N} \sum_{m=1}^{N} \mathrm{E}\left(X_n X_m^*\right) - \left|\mathrm{E}\left(X\right)\right|^2$$

$$\mathrm{E}\left(X_n X_m\right) = \begin{cases} \mathrm{E}\left(X^2\right) & , \ n = m \\ \mathrm{E}^2\left(X\right) & , \ n \neq m \end{cases}$$

Sample Mean and Variance

In $\sum_{n=1}^{N}\sum_{m=1}^{N}\mathrm{E}\left(X_n X_m^*\right)$ there are exactly N^2 terms, N terms in which $n = m$ and in all the rest $n \neq m$. Therefore

$$\sigma_{\bar{X}}^2 = \frac{1}{N^2}\left[\sum_{n=1}^{N}\mathrm{E}\left(\left|X\right|^2\right) + \sum_{\substack{n=1\\n\neq m}}^{N}\sum_{m=1}^{N}\mathrm{E}\left(X_n\right)\mathrm{E}\left(X_m^*\right)\right] - \left|\mathrm{E}\left(X\right)\right|^2$$

$$\sigma_{\bar{X}}^2 = \frac{1}{N^2}\left[N\,\mathrm{E}\left(\left|X\right|^2\right) + N\left(N-1\right)\mathrm{E}\left(X_n\right)\mathrm{E}\left(X_m\right)^*\right] - \left|\mathrm{E}\left(X\right)\right|^2$$

Simplifying, we find that the variance of the sample mean of a random variable is the variance of the random variable itself, divided by the sample size.

$$\sigma_{\bar{X}}^2 = \sigma_X^2 / N$$

Sample Mean and Variance

The symbol commonly used for the sample variance is S_X^2 to distinguish it from the population variance σ_X^2. A natural definition for it would be

$$S_X^2 = \frac{1}{N}\sum_{n=1}^{N}\left[X_n - \mathrm{E}(X)\right]\left[X_n - \mathrm{E}(X)\right]^*$$

The expected value of this sample variance is the population variance and it is, therefore, unbiased. The problem with this definition of sample variance is that in a typical data-analysis situation the population's expected value $\mathrm{E}(X)$ is probably unknown.

Sample Mean and Variance

Since the sample mean is known and it is an unbiased estimator of the population mean we could re-define the sample variance as

$$S_X^2 = \frac{1}{N}\sum_{n=1}^{N}\left(X_n - \bar{X}\right)\left(X_n - \bar{X}\right)^*$$

The expected value of this sample variance is

$$\mathrm{E}\left(S_X^2\right) = \frac{N-1}{N}\sigma_X^2$$

Therefore this is a **biased** estimator.

Sample Mean and Variance

The sample variance can be defined in such a way as to make it unbiased. That definition is

$$S_X^2 = \frac{1}{N-1}\sum_{n=1}^{N}\left(X_n - \overline{X}\right)\left(X_n - \overline{X}\right)^* = \frac{1}{N-1}\sum_{n=1}^{N}\left|X_n - \overline{X}\right|^2$$

This will be the definition used from here on. The variance of this sample variance can be shown to be

$$\mathrm{Var}\left(S_X^2\right) = \frac{N\left[\mathrm{E}\left(\left|X - \mathrm{E}\left(X\right)\right|^4\right) - \left(\sigma_X^2\right)^2\right]}{\left(N-1\right)^2}$$

Median and Mode

There are two other commonly-used descriptors of random data, the **mode** and the **median**. The mode of a set of data is the data value that occurs most often. If there are multiple data values that all occur the same number of times and all other values occur less often, the set of data is said to be **multimodal**.

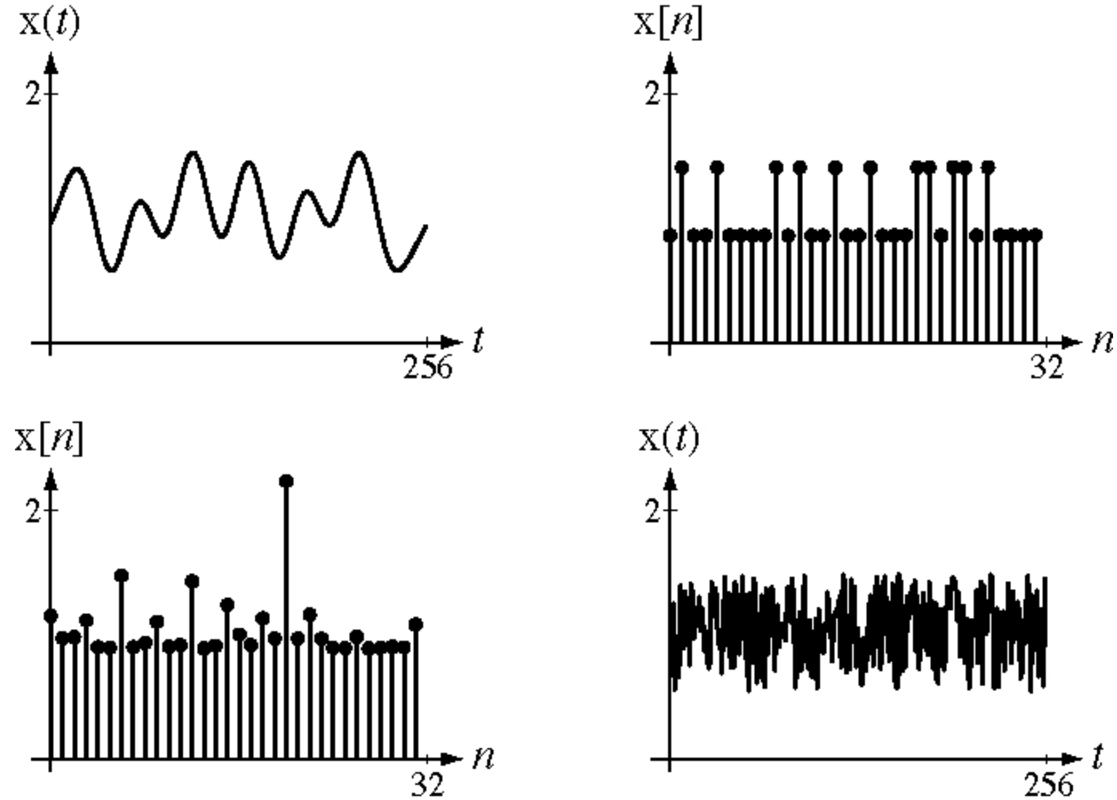
$$\mathrm{P}\left[x_{mode}\right] \geq \mathrm{P}\left[X\right]$$

The median of a set of data is the value for which an equal number of the data values fall above and below it.

$$\mathrm{P}\left[X > x_{median}\right] = \mathrm{P}\left[X < x_{median}\right]$$

Histograms and Probability Density

The four signals illustrated all have the same mean and variance. Another descriptor that distinguishes them from each other is a **histogram**. A histogram is a plot of the number of times each data value occurs in a sample versus those values.



Histograms and Probability Density

Suppose the data collected from 40 trials of an experiment are

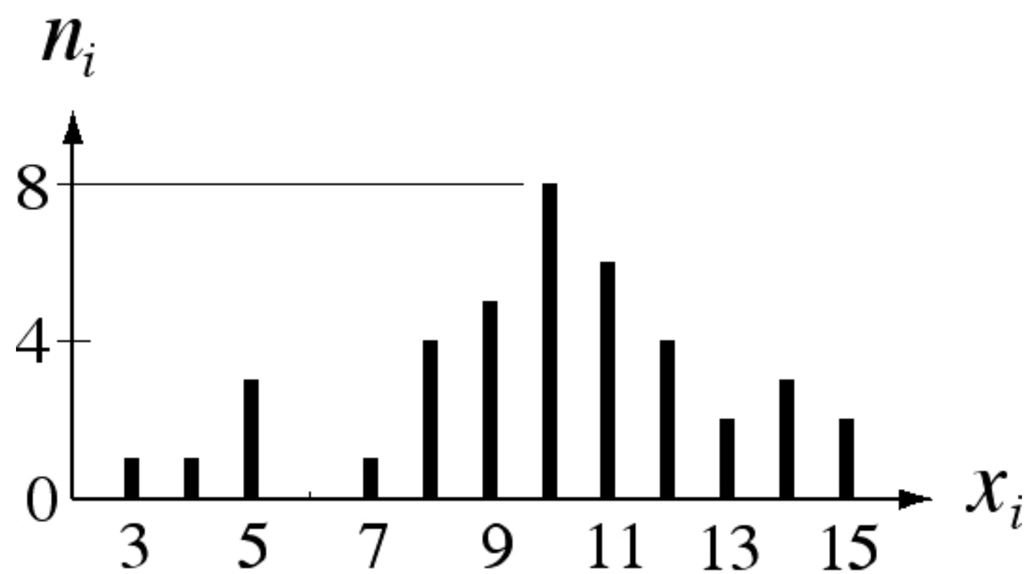
$$\begin{Bmatrix} 10, & 14, & 10, & 12, & 13, & 11, & 8, & 9, & 7, & 3, \\ 10, & 11, & 10, & 9, & 8, & 13, & 11, & 14, & 12, & 14, \\ 5, & 9, & 12, & 15, & 4, & 10, & 15, & 10, & 8, & 5, \\ 9, & 11, & 10, & 9, & 10, & 11, & 12, & 5, & 11, & 8 \end{Bmatrix}$$

One way to better understand the data is to tabulate them $\longrightarrow$

x_i	n_i	$\mathrm{r}(x_i)$	$x_i\,\mathrm{r}(x_i)$
3	1	0.025	0.075
4	1	0.025	0.1
5	3	0.075	0.375
6	0	0	0
7	1	0.025	0.175
8	4	0.1	0.8
9	5	0.125	1.125
10	8	0.2	2
11	6	0.15	1.65
12	4	0.1	1.2
13	2	0.05	0.65
14	3	0.075	1.05
15	2	0.05	0.75
	Totals	1	9.95

Histograms and Probability Density

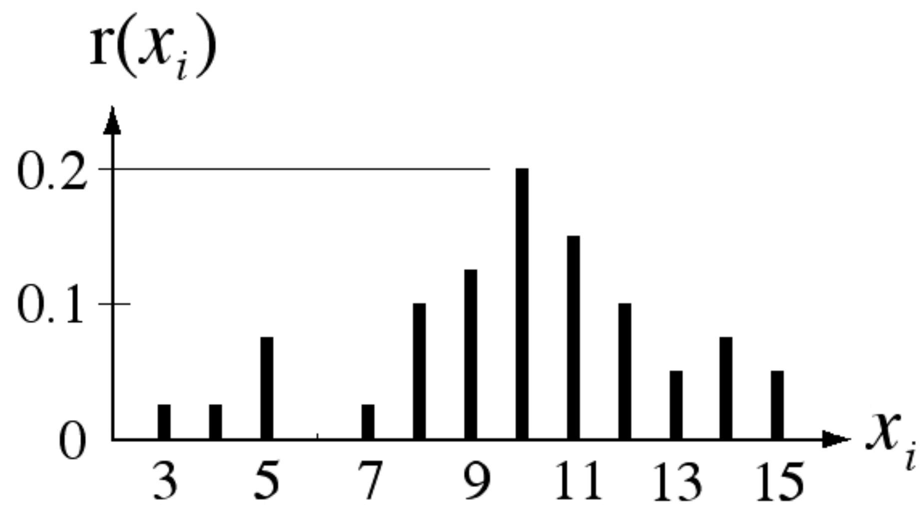
To aid understanding of the general nature of the random variable, the data can be plotted as a histogram.



It is now obvious that there is a **central tendency** in the data.

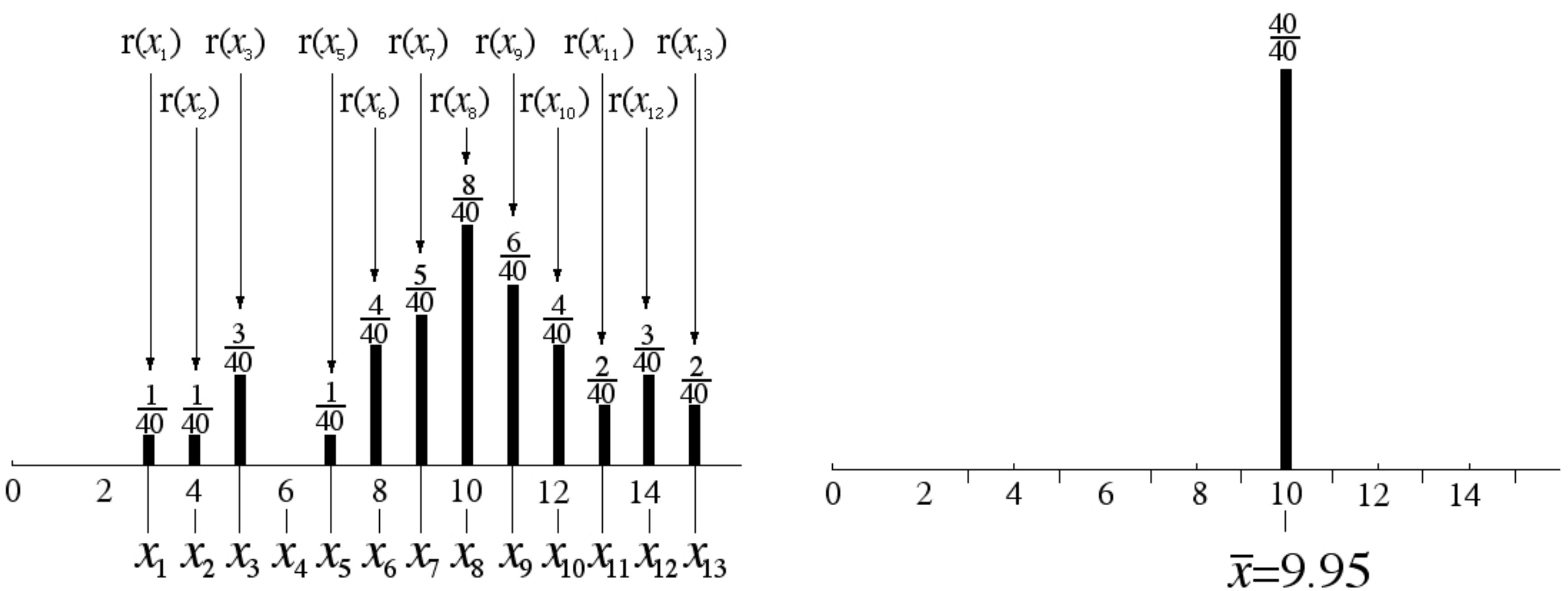
Histograms and Probability Density

As larger samples of data are taken the accuracy of the histogram in representing the general distribution of the data values improves. In the limit as the sample size approaches infinity the histogram becomes perfect but it cannot be plotted because all the numbers of occurrences are infinite. It is often better to plot a relative-frequency histogram.



Histograms and Probability Density

An analogy in classical mechanics is useful in conceiving mean value. The values are the moment-arm lengths and the relative frequencies of occurrence are the weights. The same moment occurs with the mean as the moment-arm length and a weight of one.

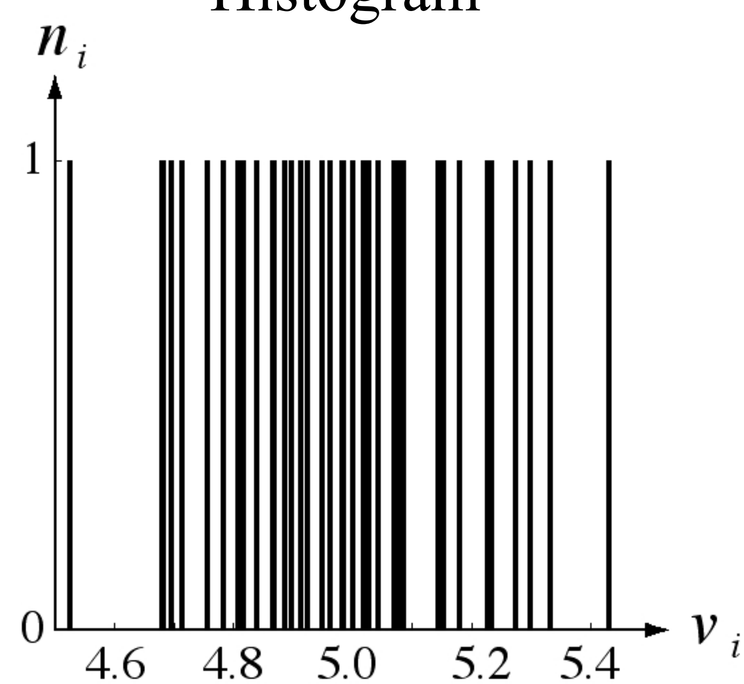


Histograms and Probability Density

Suppose the data from multiple trials of another experiment are

4.8852	5.2726	5.2297	4.9242
4.9119	4.6955	5.3325	4.9488
4.8681	5.0771	4.9625	4.6803
5.1521	4.8168	5.2981	5.4297
4.8972	4.9827	4.8663	5.0428
4.5239	4.9837	4.7553	4.7832
4.8385	4.9845	4.6809	5.2333
4.8079	5.0261	5.1787	5.0697
5.2260	4.7131	5.0856	5.0715
5.0200	5.1443	5.0183	4.9995

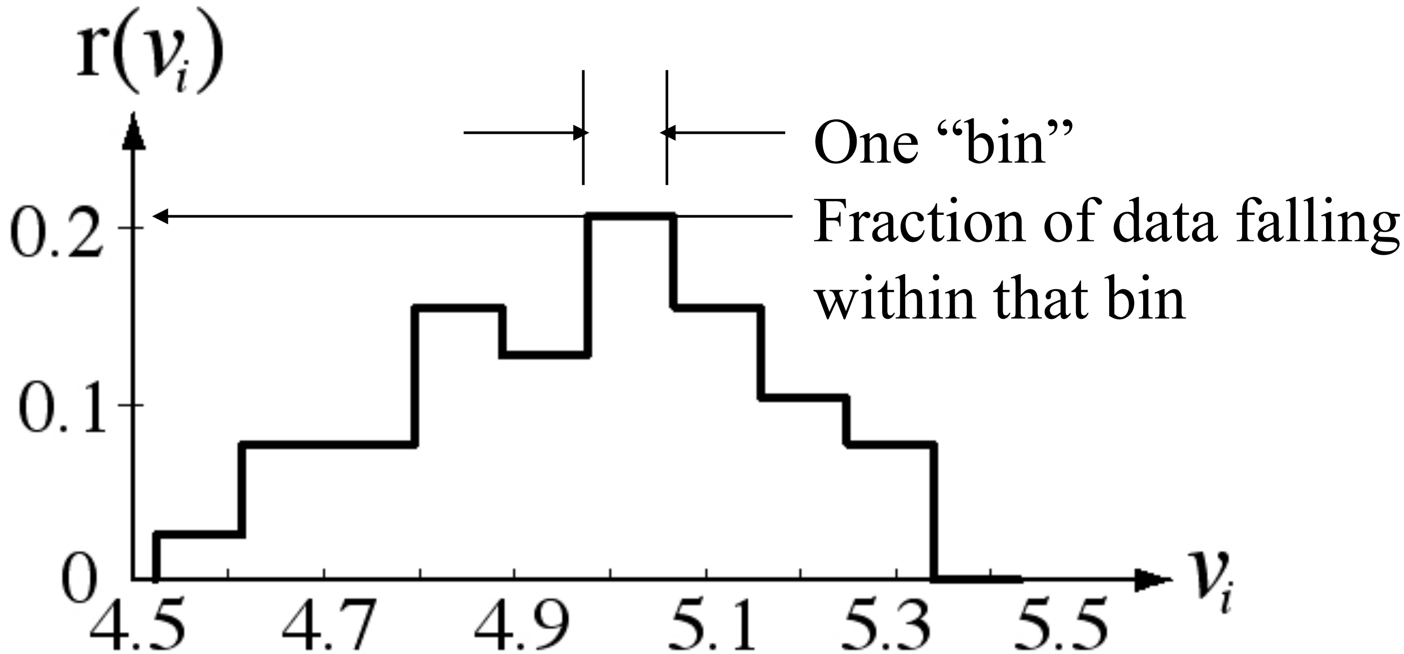
Histogram



In this case, the histogram is not quite as useful

Histograms and Probability Density

An alternate form of histogram is better for data in which there are few, if any, repetitions.

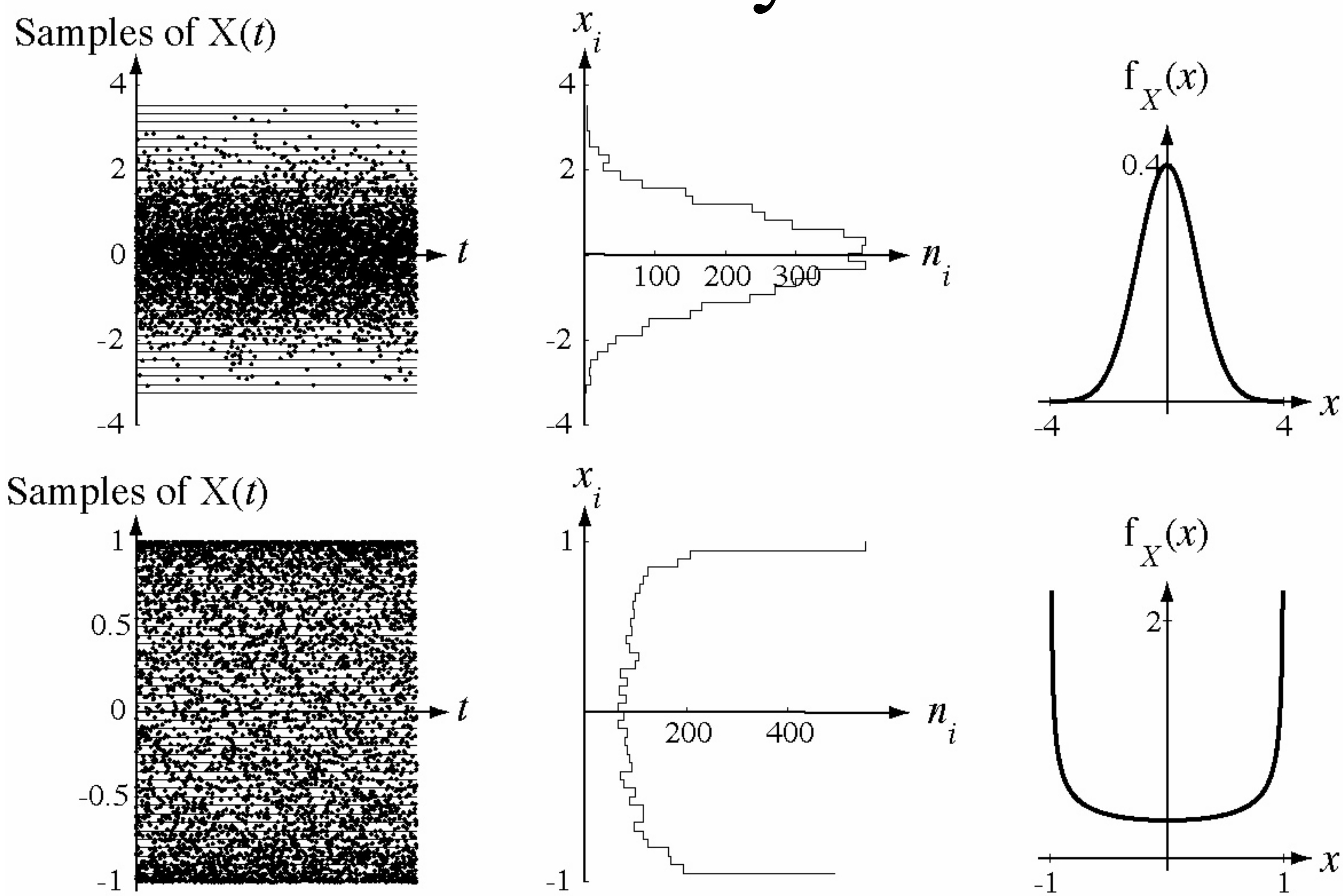


Histograms and Probability Density

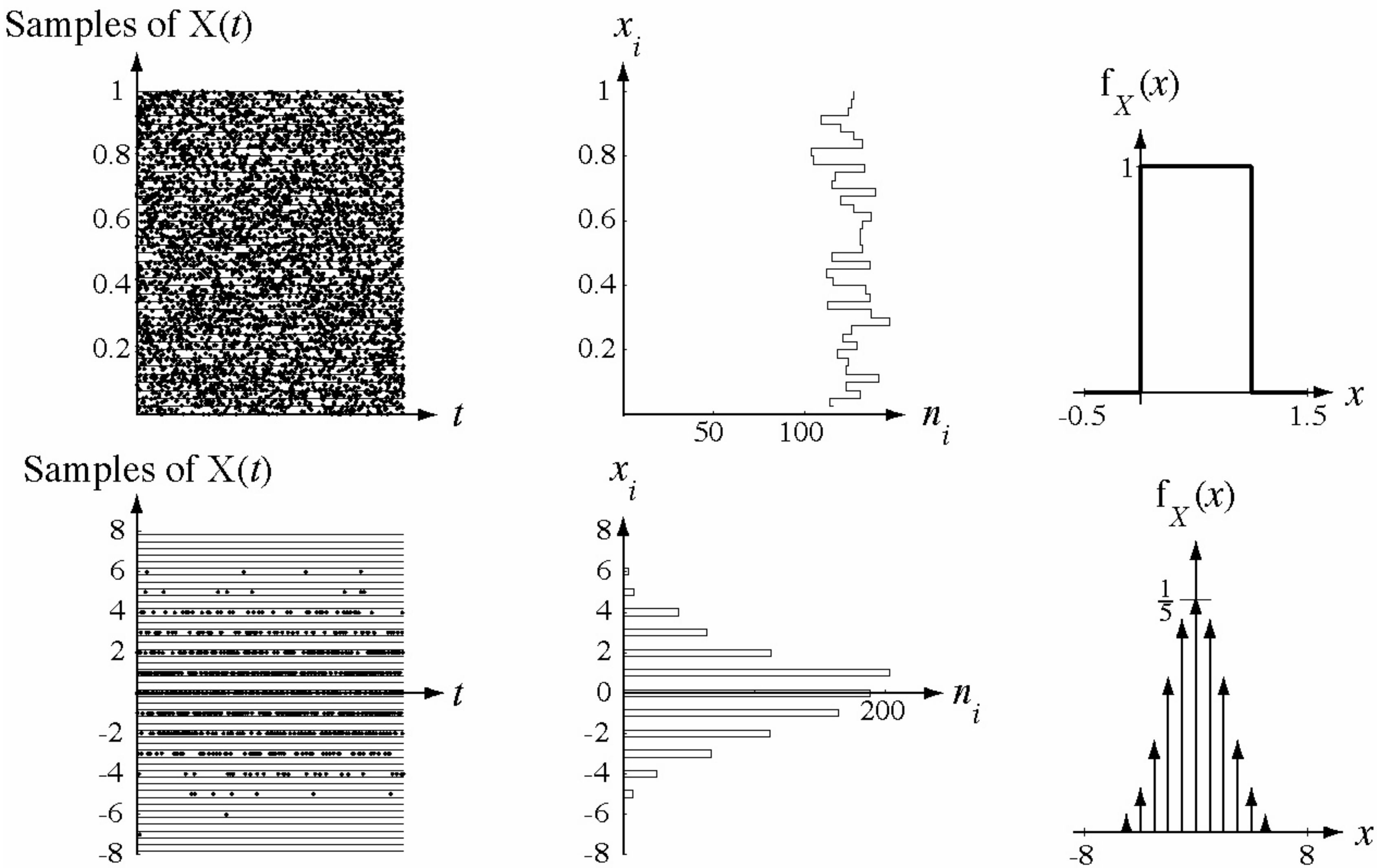
One way of conceiving the probability density function is as the limit of a relative frequency histogram normalized to the bin width as the number of bins approaches infinity

$$\mathrm{f}_X(x) = \lim_{\substack{N\to\infty \\ \Delta x\to 0}} \frac{n}{N\Delta x} = \lim_{\substack{N\to\infty \\ \Delta x\to 0}} \frac{\mathrm{r}(x)}{\Delta x}$$

Histograms and Probability Density



Histograms and Probability Density



Maximum Likelihood Estimation

Let $\mathbf{x}_n = \{x_1, x_2, \cdots, x_n\}$ be the observed values of a random sampling of the random variable X and let θ be the parameter we want to estimate. The **likelihood function** is

$$\ell(\mathbf{x}_n;\theta) = \ell(x_1, x_2, \cdots, x_n;\theta) = \begin{cases} \mathrm{p}_X(x_1, x_2, \cdots, x_n \mid \theta) & \text{if } X \text{ is a DV random variable} \\ \mathrm{f}_X(x_1, x_2, \cdots, x_n \mid \theta) & \text{if } X \text{ is a CV random variable} \end{cases}$$

Since the samples are iid

$$\mathrm{p}_X(x_1, x_2, \cdots, x_n \mid \theta) = \mathrm{p}_X(x_1 \mid \theta)\mathrm{p}_X(x_2 \mid \theta)\cdots\mathrm{p}_X(x_n \mid \theta) = \prod_{j=1}^{n} \mathrm{p}_X(x_j \mid \theta)$$

and

$$\mathrm{f}_X(x_1, x_2, \cdots, x_n \mid \theta) = \mathrm{f}_X(x_1 \mid \theta)\mathrm{f}_X(x_2 \mid \theta)\cdots\mathrm{f}_X(x_n \mid \theta) = \prod_{j=1}^{n} \mathrm{f}_X(x_j \mid \theta)$$

Maximum Likelihood Estimation

The **maximum likelihood method** selects the estimator value $\hat{\Theta} = \theta^*$ where θ^* is the parameter value that maximizes the likelihood function.

$$\ell\left(x_1, x_2, \cdots, x_n; \theta^*\right) = \max_{\theta} \ell\left(x_1, x_2, \cdots, x_n; \theta\right)$$

It is often more convenient to work with the **log likelihood function** $\mathrm{L}\left(\mathbf{x}_n \mid \theta\right) = \ln\left(\ell\left(\mathbf{x}_n; \theta\right)\right)$ because then the iterated product becomes an iterated sum

$$\mathrm{L}\left(\mathbf{x}_n \mid \theta\right) = \sum_{j=1}^{n} \ln\left(\mathrm{p}_X\left(x_j \mid \theta\right)\right) \quad \text{or} \quad \mathrm{L}\left(\mathbf{x}_n \mid \theta\right) = \sum_{j=1}^{n} \ln\left(\mathrm{f}_X\left(x_j \mid \theta\right)\right)$$

Maximizing the log likelihood function is typically done by finding the value θ^* for which $\dfrac{\partial}{\partial \theta} \mathrm{L}\left(\mathbf{x}_n \mid \theta\right) = 0$.

Maximum Likelihood Estimation

Example:

Let $\mathbf{x}_n = \{x_1, x_2, \cdots, x_n\}$ be iid samples from a gamma random variable X

with unknown parameters α and λ and PDF $f_X(x) = \dfrac{\lambda(\lambda x)^{\alpha-1} e^{-\lambda x}}{\Gamma(\alpha)}$, $x > 0$, $\alpha > 0$, $\lambda > 0$.

Then $\ell(x_1, x_2, \cdots, x_n; \alpha, \lambda) = \prod_{m=1}^{n} \dfrac{\lambda(\lambda x)^{\alpha-1} e^{-\lambda x_m}}{\Gamma(\alpha)} x_m^{\alpha-1} = \dfrac{\lambda^{n\alpha}}{\Gamma^n(\alpha)} e^{-\lambda \sum_{k=1}^{n} x_k} \prod_{m=1}^{n} x_m^{\alpha-1}$.

Then $\mathrm{L}(x_1, x_2, \cdots, x_n; \alpha, \lambda) = n\alpha \ln(\lambda) - n\ln(\Gamma(\alpha)) - \lambda \sum_{k=1}^{n} x_k + (\alpha - 1)\sum_{m=1}^{n} \ln(x_m)$

Differentiating w.r.t. α and λ, and setting the derivatives equal to zero,

$$n\ln\left(\hat{\lambda}\right) - n\underbrace{\frac{\Gamma'(\hat{\alpha})}{\Gamma(\hat{\alpha})}}_{=\Psi(\alpha)} + \sum_{m=1}^{n} \ln(x_m) = 0 \text{ and } \frac{n\hat{\alpha}}{\hat{\lambda}} - \sum_{k=1}^{n} x_k = 0$$

Maximum Likelihood Estimation

Solving, $\hat{\lambda} = \dfrac{\hat{\alpha}}{\frac{1}{n}\sum_{k=1}^{n} x_k}$ and, substituting this into the first equation above

$$n \ln\left(\frac{\hat{\alpha}}{\frac{1}{n}\sum_{k=1}^{n} x_k}\right) - n\Psi(\alpha) + \sum_{k=1}^{n} \ln(x_k) = 0$$

$$\ln(\hat{\alpha}) - \Psi(\alpha) = \ln\left(\frac{1}{n}\sum_{k=1}^{n} x_k\right) - \frac{1}{n}\sum_{k=1}^{n} \ln(x_k)$$

This equation is nonlinear so solving for $\hat{\alpha}$ can only be done numerically.

Sampling Distributions

The sample mean is $\bar{X} = \frac{1}{N}\sum_{i=1}^{N} X_i$. If the samples are independent and come from a Gaussian population then the random variable

$$Z = \frac{\overline{X} - \mathrm{E}(X)}{\sigma_X / \sqrt{N}}$$

is normally distributed. For large N the population variance may be replaced by the sample variance with negligible error.

Sampling Distributions

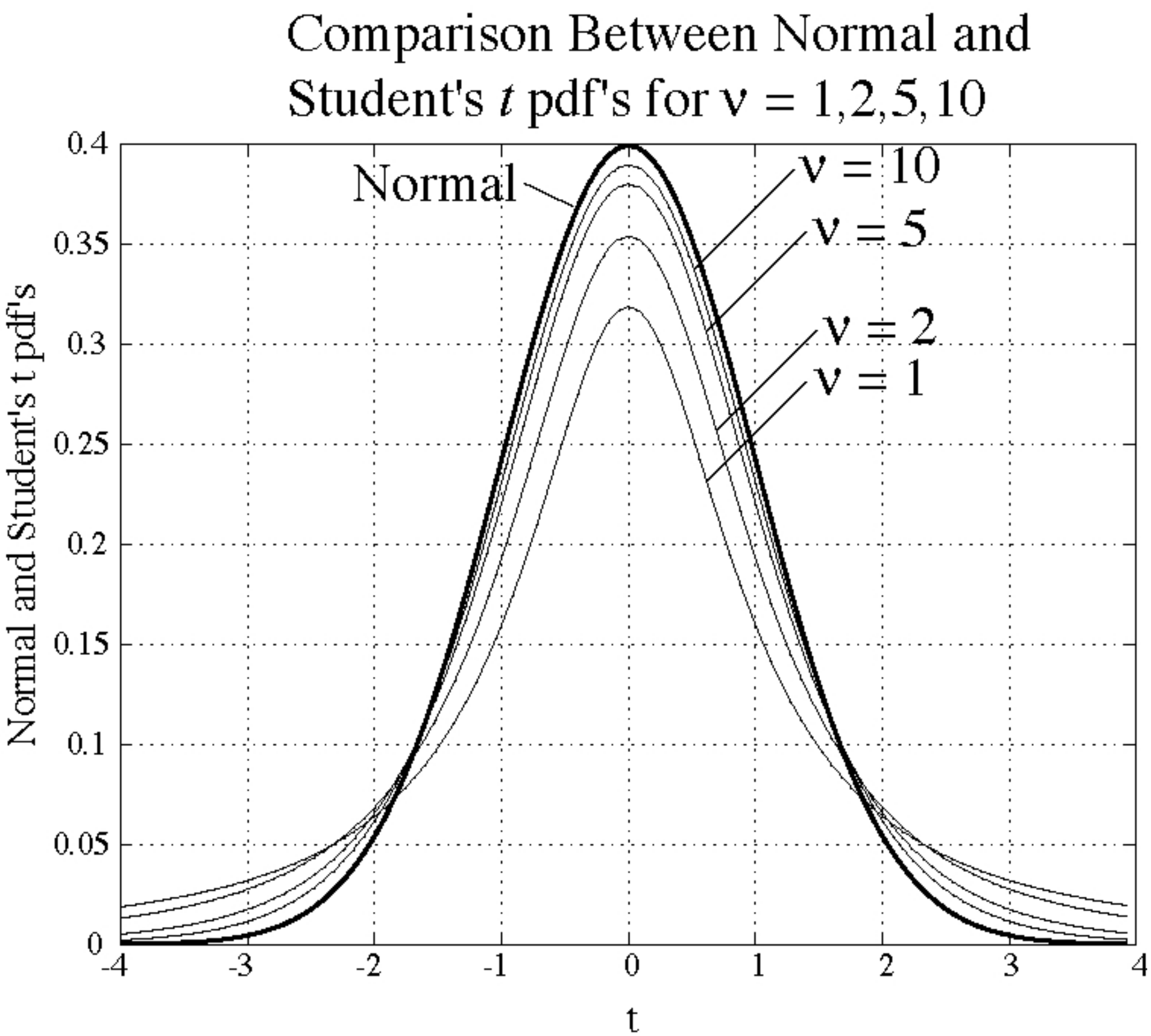
For small N the sample variance is not as good an estimate of the population variance. Define a new random variable,

$$T = \frac{\bar{X} - \mathrm{E}(X)}{S_X / \sqrt{N}}$$

Since the sample variance S_{X}^2 is not a constant but rather a random variable, the variance of T is larger than the variance of Z. The PDF of T was found by William Gosset and is called the "Student's t distribution" PDF.

$$\mathrm{p}_T(t) = \frac{\Gamma\left(\left(\nu+1\right)/2\right)}{\sqrt{\nu\pi}\,\Gamma\left(\nu/2\right)}\left(1+t^2/\nu\right)^{-\left(\nu+1\right)/2}$$

Sampling Distributions



Sampling Distributions

Statistical results are often reported in terms of **confidence intervals** and **confidence levels**. A confidence interval is a range in which a random variable can be expected to lie, with a corresponding confidence level indicating the probability that the random variable lies in that interval. For any given random variable, as the confidence interval is increased, the confidence level increases.

A Gaussian distributed random variable may be expected to lie within some multiple of its standard deviation from the mean, with a level of confidence determined by the Gaussian PDF.

Sampling Distributions

For a Gaussian-distributed random variable,

Confidence Interval	Confidence Level
$\pm\sigma_{\bar{X}}$	68.3%
$\pm 1.64\sigma_{\bar{X}}$	90%
$\pm 1.96\sigma_{\bar{X}}$	95%
$\pm 2\sigma_{\bar{X}}$	95.45%
$\pm 2.58\sigma_{\bar{X}}$	99%
$\pm 3\sigma_{\bar{X}}$	99.73%
$\pm 3.29\sigma_{\bar{X}}$	99.9%
$\pm 3.89\sigma_{\bar{X}}$	99.99%

Hypothesis Testing

Hypothesis testing is a process of deciding between two alternatives for a parameter of a population based on a sample estimate of that parameter from that population. There are two alternatives, the **null hypothesis** H_0 and the **alternative hypothesis** H_1. The null hypothesis is usually that a population parameter has a certain value. The alternative hypothesis may be that the parameter does not have that value or that the parameter has a value less than that value or a value more than that value. For example he null hypothesis might be that the mean of a population is 20 and the alternative hypothesis might be that the mean of the population is not 20, usually written as

$$H_0: \ \mu = 20 \ , \ H_1: \ \mu \neq 20$$

Hypothesis Testing

$$H_0: \ \mu = 20 \ \ , \ H_1: \ \mu \neq 20$$

This type of alternative hypothesis is called **two - sided** because it can be satisfied by a value either greater than or less than the null hypothesis value. The hypotheses might instead be **one - sided**.

$H_0: \ \mu = 20 \ \ , \ H_1: \ \mu < 20$ or $H_0: \ \mu = 20 \ \ , \ H_1: \ \mu > 20$

The actual process of making the decision between two alternatives is called a **test of the hypothesis**. The null hypothesis is accepted if the estimate of the parameter based on a sample is consistent with the null hypothesis. Being consistent means that the sample parameter is within the **acceptance region**. All other values of the sample parameter are within the **critical region**.

Hypothesis Testing

If the sample parameter falls within the acceptance region we accept the null hypothesis. Otherwise we reject the null hypothesis. Rejecting the null hypothesis when it should be accepted is called a **Type I Error**. Accepting the null hypothesis when it should be rejected is called a **Type II Error**. The probability of making a Type I Error is conventionally designated by α and the probability of making a Type II Error is designated by β. α is also sometimes called the **significance level** of the test of the hypothesis.

Hypothesis Testing

Example

Suppose the actual population mean LED optical power of LED's made in a certain manufacturing process is 5 mW and the population standard deviation of LED optical power is 0.5 mW and that the population has a Gaussian pdf. Further, let the null hypothesis be that the population mean is 5 mW and let the acceptance region be the range 4.9 mW to 5.1 mW. Let the alternative hypothesis be that the population mean is not 5 mW, making it two-sided. Suppose the sample size is 80. What is the probability of rejecting the null hypothesis?

Hypothesis Testing

Example

The standard deviation of the sample mean is

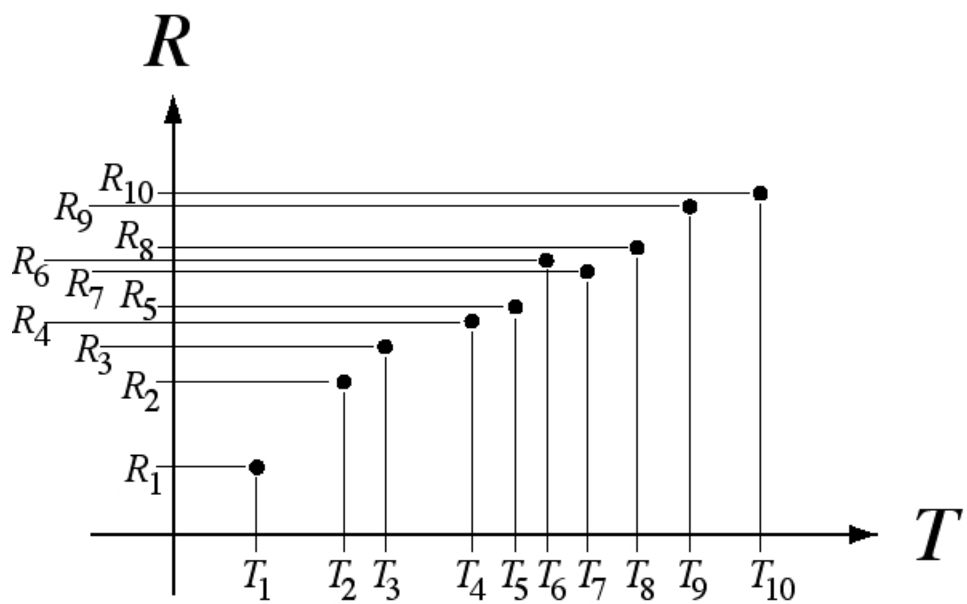
$$\sigma_{\bar{X}} = \frac{\sigma_X}{\sqrt{80}} = \frac{0.5 \text{ mW}}{8.944} = 0.0559 \text{ mW}$$

The probability of rejecting the null hypothesis is

$$\alpha = \mathrm{G}\left(\frac{4.9-5}{0.0559}\right) + 1 - \mathrm{G}\left(\frac{5.1-5}{0.0559}\right) = 0.0368 + 0.0368 = 0.0736$$

Curve Fitting

A calibration of a platinum resistance thermometer might look like this where *R* is resistance and *T* is temperature.

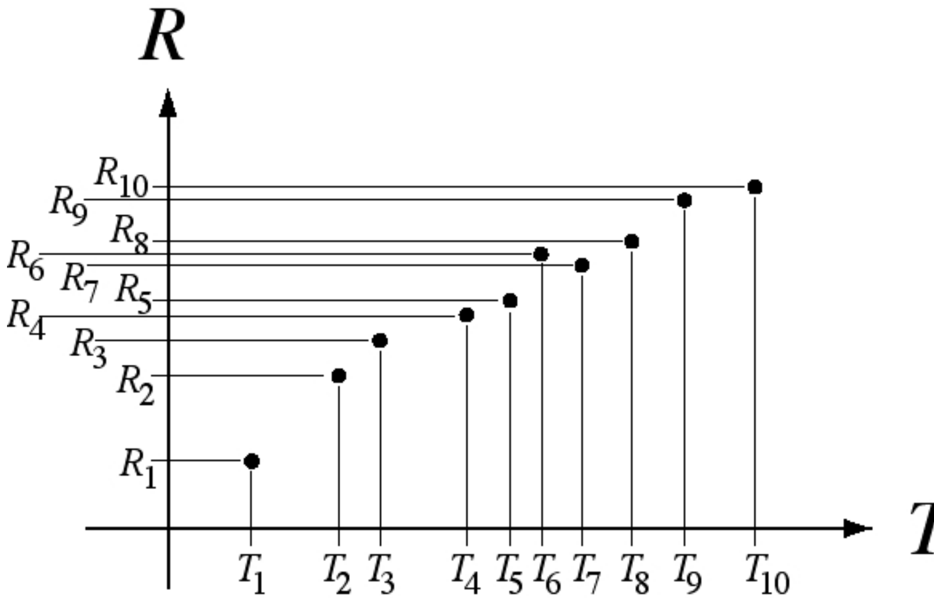


Since we know platinum's resistance does not actually vary exactly this way what is the best interpretation of the data?

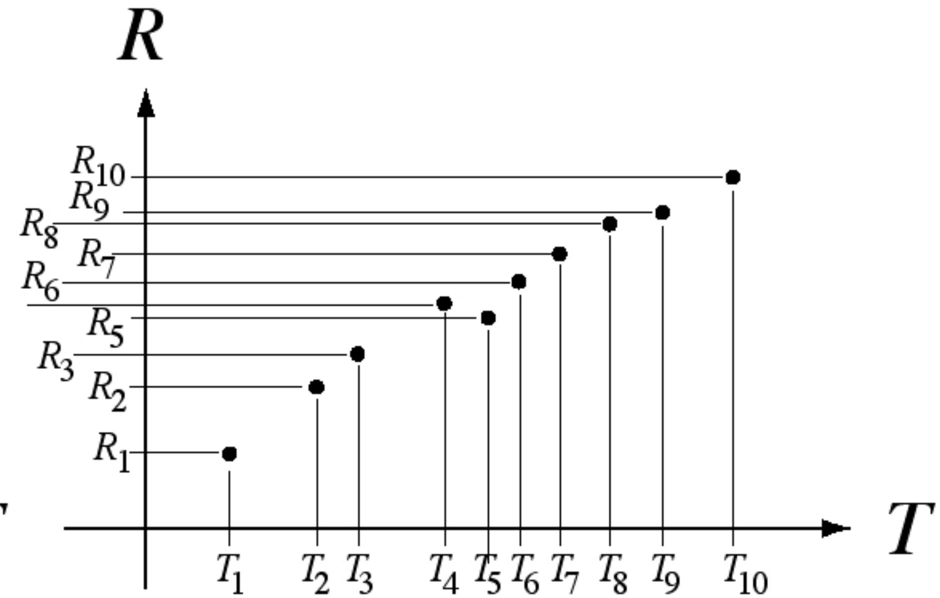
Curve Fitting

A repeated calibration might look like the second graph below

1st Calibration



2nd Calibration



The general trend is the same but the details are different. The differences are caused by measurement error.

Curve Fitting

The best interpretation of the calibration data is that the relation between R and T should be "smooth" and "slowly-varying" and it should ignore the small fluctuations caused by measurement errors. At the same time the mean-squared error between the R - T relation and the data should be minimized.

The simplest way to satisfy these criteria is to find a straight line relation between R and T with these qualities. That is, we want to find a function of the form, $\mathrm{R}(T) = a_0 + a_1 T$ where the two a coefficients are chosen to minimize the mean-squared error.

Curve Fitting

The relation between the calibration data and the best-fit line can be expressed as

$$R_1 = a_0 + a_1 T_1 + \varepsilon_1$$
$$R_2 = a_0 + a_1 T_2 + \varepsilon_2$$
$$\vdots$$
$$R_N = a_0 + a_1 T_N + \varepsilon_N$$

where N is the number of measurements and the ε's represent the random measurement error. Then the sum-squared error SSE is

$$SSE = \sum_{i=1}^{N} \varepsilon_i^2 = \sum_{i=1}^{N} \left(R_i - a_0 - a_1 T_i \right)^2$$

Curve Fitting

Setting the derivatives with respect to the two a's to zero we get

$$\frac{\partial\left(SSE\right)}{a_0} = -2\sum_{i=1}^{N}\left(R_i - \hat{a}_0 - \hat{a}_1 T_i\right) = 0$$

$$\frac{\partial\left(SSE\right)}{a_1} = -2\sum_{i=1}^{N}T_i\left(R_i - \hat{a}_0 - \hat{a}_1 T_i\right) = 0$$

where the little "hats" on the a's indicate that we will find estimates of the "actual" a's because we are basing the estimates on a finite set of data. The **normal** equations can then be written as

$$N\hat{a}_0 + \hat{a}_1\sum_{i=1}^{N}T_i = \sum_{i=1}^{N}R_i$$

$$\hat{a}_0\sum_{i=1}^{N}T_i + \hat{a}_1\sum_{i=1}^{N}T_i^2 = \sum_{i=1}^{N}T_i R_i$$

Curve Fitting

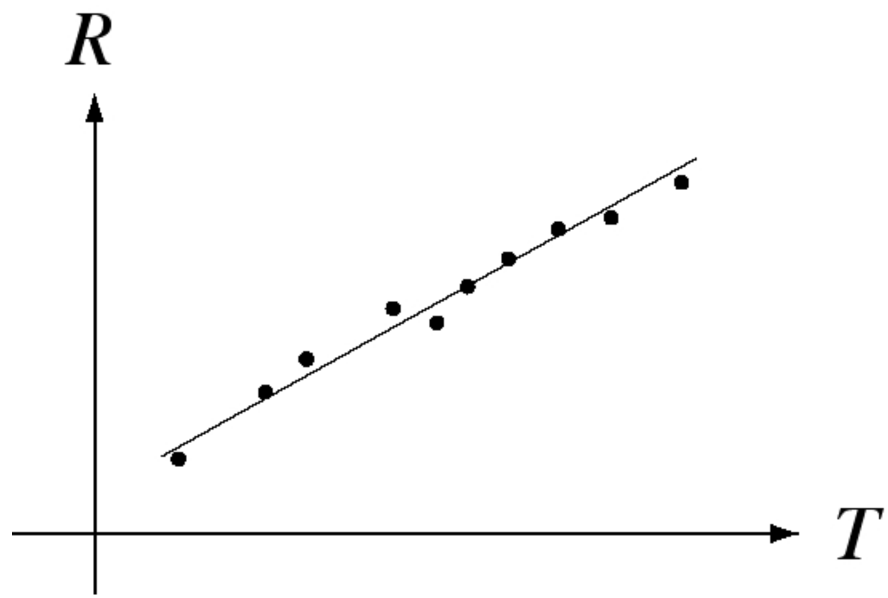
The solutions of the normal equations are

$$\hat{a}_0 = \frac{\left(\sum_{i=1}^{N} R_i\right)\left(\sum_{i=1}^{N} T_i^2\right) - \left(\sum_{i=1}^{N} T_i R_i\right)\left(\sum_{i=1}^{N} T_i\right)}{N\sum_{i=1}^{N} T_i^2 - \left(\sum_{i=1}^{N} T_i\right)^2}$$

$$\hat{a}_1 = \frac{N\sum_{i=1}^{N} T_i R_i - \left(\sum_{i=1}^{N} T_i\right)\left(\sum_{i=1}^{N} R_i\right)}{N\sum_{i=1}^{N} T_i^2 - \left(\sum_{i=1}^{N} T_i\right)^2}$$

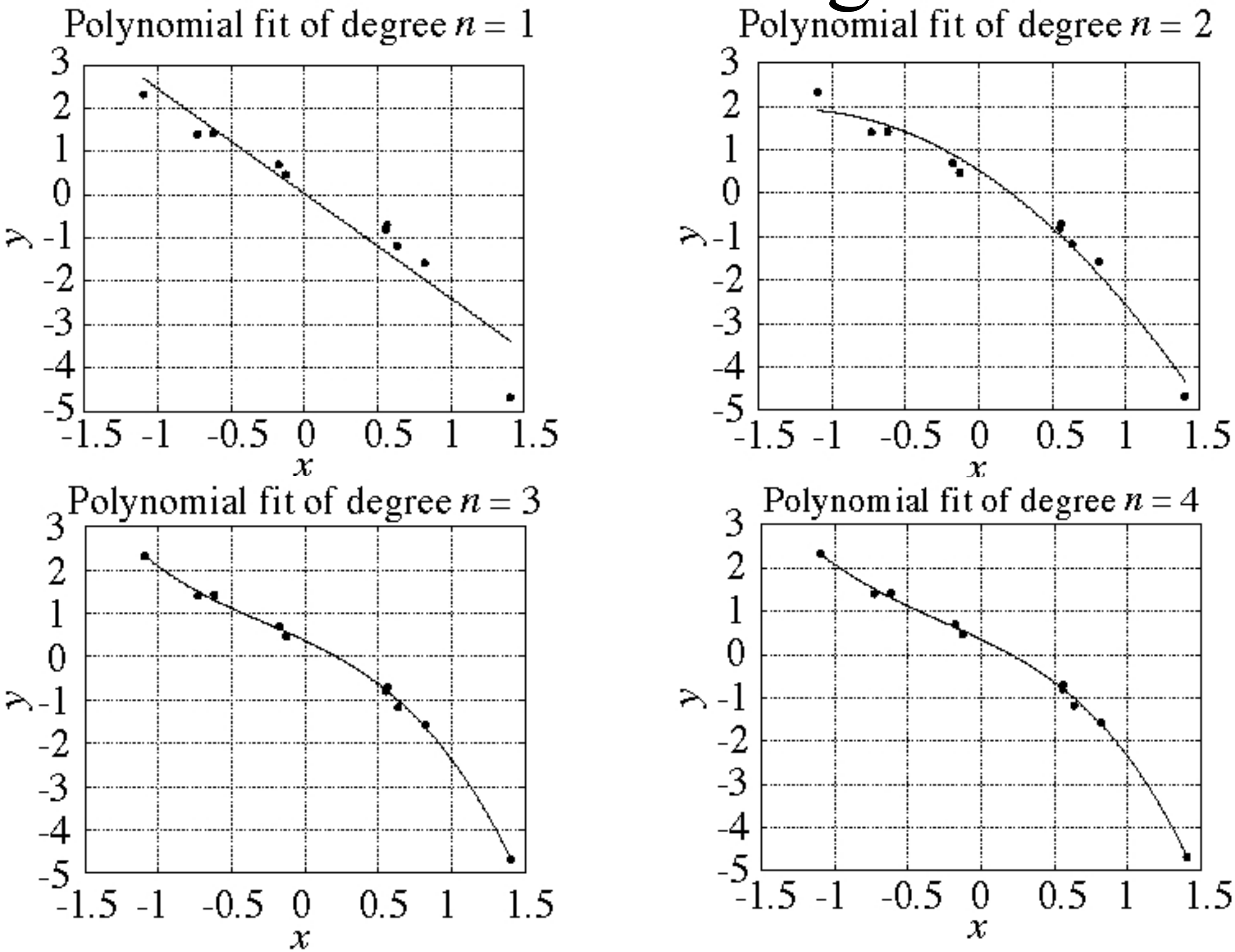
Curve Fitting

The best-fit straight line might look like this

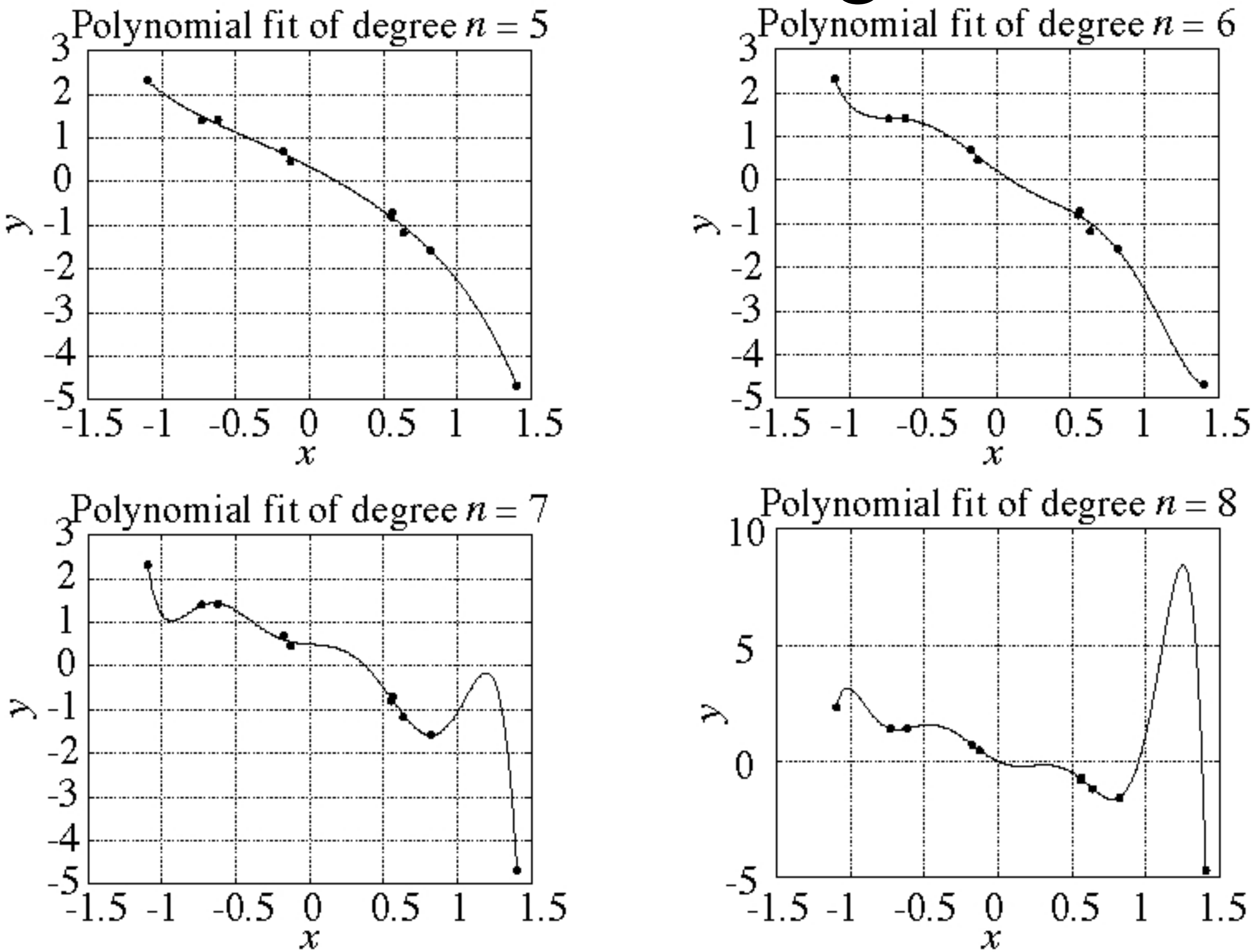


In observing this straight-line fit, it may seem that the actual relationship might be curved rather than a straight line. The fitting function could be a second order (or higher order) polynomial instead.

Curve Fitting



Curve Fitting



Multiple Linear Regression

The most general form of curve fitting is **multiple linear regression**. This technique assumes that a variable Y is a function of multiple other variables X.

$$Y\left(X_1, X_2, \ldots, X_k\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k + \varepsilon$$

The measurements on the measured variable Y are then

$$\begin{aligned}
y_1 &= \beta_0 + \beta_1 x_{11} + \beta_2 x_{12} + \cdots + \beta_k x_{1k} + \varepsilon_1 \\
y_2 &= \beta_0 + \beta_1 x_{21} + \beta_2 x_{22} + \cdots + \beta_k x_{2k} + \varepsilon_2 \\
&\ \ \vdots \qquad\quad \vdots \qquad\quad \vdots \\
y_N &= \beta_0 + \beta_1 x_{N1} + \beta_2 x_{N2} + \cdots + \beta_k x_{Nk} + \varepsilon_N
\end{aligned}$$

Multiple Linear Regression

To minimize the sum-squared error,

$$SSE\left(\beta_0,\beta_1,\beta_2,\cdots\beta_k\right)=\sum_{i=1}^{N}\left(y_i-\beta_0-\sum_{\ell=1}^{k}\beta_\ell x_{i\ell}\right)^2$$

differentiate with respect to the β's and set the derivatives equal to zero.

$$\frac{\partial SSE}{\partial\beta_j}=\sum_{i=1}^{N}\left[-2x_{ij}\left(y_i-\beta_0-\sum_{\ell=1}^{k}\beta_\ell x_{i\ell}\right)\right]\ ,\ \left(x_{i0}\equiv 1\right)$$

Multiple Linear Regression

The mathematics of multiple linear regression can be compactly written in terms of vectors and matrices.

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix} \quad \mathbf{x}_i = \begin{bmatrix} 1 & x_{i1} & x_{i2} & \cdots & x_{ik} \end{bmatrix} \quad \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix}, \quad \varepsilon = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_N \end{bmatrix}$$

$$\mathbf{X} = \begin{bmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \\ \vdots \\ \mathbf{x}_N \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1k} \\ 1 & x_{21} & x_{22} & \cdots & x_{2k} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{N1} & x_{N2} & \cdots & x_{Nk} \end{bmatrix}$$

Multiple Linear Regression

Then, in matrix form,

$$\mathbf{y} = \mathbf{X}\beta + \varepsilon$$

The solution for the estimates of the β's is

$$\hat{\beta} = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T \mathrm{y}$$

and the best estimate of Y for a new x is

$$\hat{y} = \mathbf{x}\hat{\beta} = \hat{\beta}_0 + \sum_{\ell=1}^{k} \hat{\beta}_\ell x_\ell$$