# STATISTICS OF OBSERVATIONS & SAMPLING THEORY

**References:**

**Bevington "Data Reduction & Error Analysis for the Physical Sciences"**

**LLM: Appendix B**

**Warning: the introductory literature on statistics of measurement is remarkably uneven, and nomenclature is not consistent.**

**Is error analysis important? Yes! See next page.**

## Parent Distributions

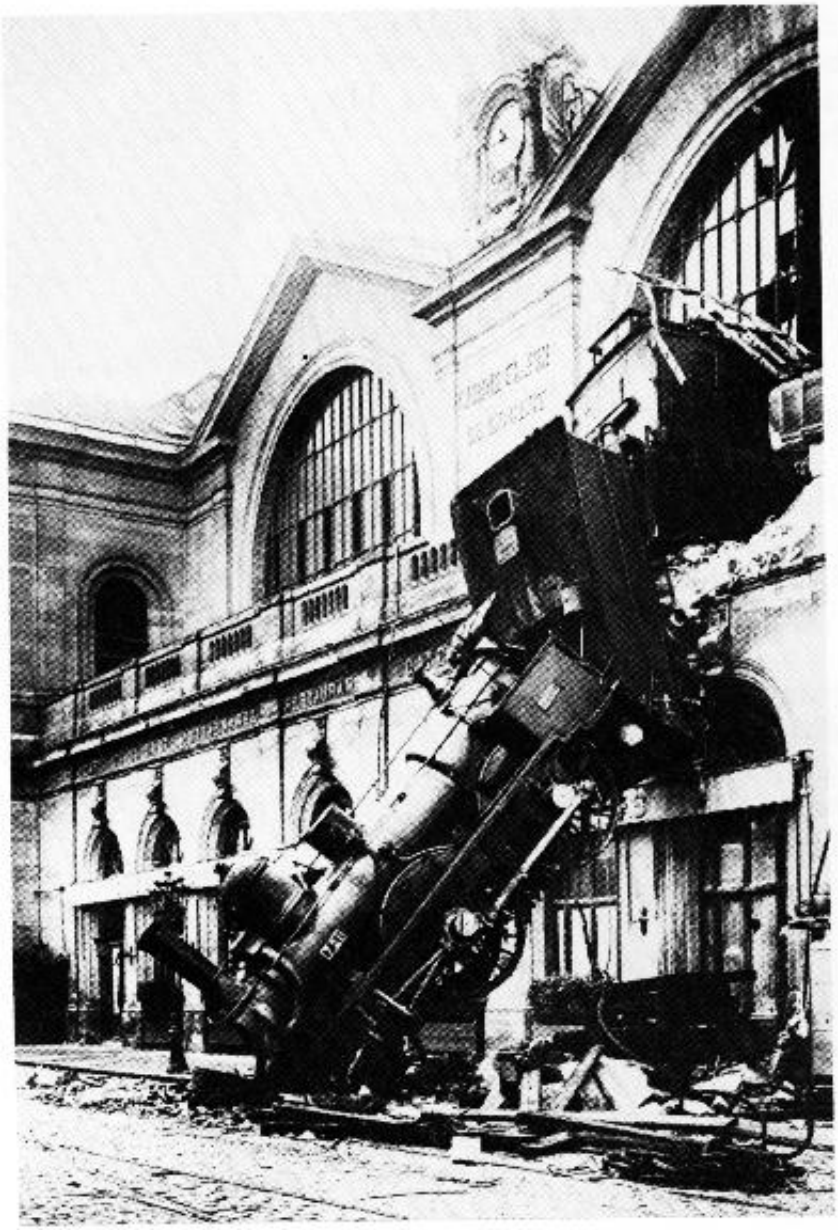
**Measurement of any physical quantity is always affected by uncontrollable random ("stochastic") processes. These produce a statistical scatter in the values measured.**

**The parent distribution for a given measurement gives the probability of obtaining a particular result from a single measure. It is fully defined and represents the idealized outcome of an infinite number of measures, where the random effects on the measuring process are assumed to be always the same ("stationary").**

# An Introduction to Error Analysis

## *The Study of Uncertainties in Physical Measurements*

John R. Taylor



# Precision vs. Accuracy

- **The parent distribution only describes the stochastic scatter in the measuring process. It does not characterize how close the measurements are to the true value of the quantity of interest. Measures can be affected by systematic errors as well as by random errors.**

  **In general, the effects of systematic errors are not manifested as stochastic variations during an experiment. In the lab, for instance, a voltmeter may be improperly calibrated, leading to a bias in all the measurements made. Examples of potential systematic effects in astronomical photometry include a wavelength mismatch in CCD flat-field calibrations, large differential refraction in Earth's atmosphere, or secular changes in thin clouds.**

- **Distinction between precision and accuracy:**
  - **A measurement with a large ratio of value to statistical uncertainty is said to be "precise."**
  - **An "accurate" measurement is one which is close to the true value of the parameter being measured.**
  - **Because of systematic errors precise measures may not be accurate.**
  - **A famous example: the primary mirror for the Hubble Space Telescope was figured with high precision (i.e. had very small ripples), but it was inaccurate in that its shape was wrong.**
- **The statistical infrastructure we are discussing here does not permit an assessment of systematic errors. Those must be addressed by other means.**

# Moments of Parent Distribution

**The parent distribution is characterized by its moments:**

- **Parent probability distribution:** $p(x)$
- **Mean: first moment.** $\mu \equiv \int x\, p(x)\, dx$
- **Variance: second moment.**

$$Var(x) \equiv \int (x-\mu)^2\, p(x)\, dx$$

- **"Sigma":** $\sigma \equiv \sqrt{Var(x)}$
- **Aliases:** $\sigma$ **is the standard deviation, but is also known as the "dispersion" or "rms dispersion"**
- $\mu$ **measures the "center" and** $\sigma$ **measures the "width" of the parent distribution.**

> **NB: the mean can be very different from the median (**$50^{th}$ **percentile) or the mode (most frequent value) of the parent distribution. These represent alternative measures of the distribution's "center." But the mean is the more widely used parameter.**

# Poisson Probability Distribution

**Applies to any continuous counting process where events are independent of one another and have a uniform probability of occurring in any time bin.**

> **The Poisson distribution is derived as a limit of the "binomial distribution" based on the fact that time can be divided up into small intervals such that the probability of an event in any given interval is arbitrarily small.**

**If $n$ is the number of counts observed in one $\delta t$ bin, then**

$$p_P(n) = \frac{\mu^n}{n!}\, e^{-\mu}$$

**Properties:**

$$\sum_{n=0}^{\infty} p_P(n) = 1$$

**Asymmetric about $\mu$; mode $\leq \mu$**

**Mean value per bin: $\mu$. $\mu$ need not be an integer.**

**Variance: $\mu$**

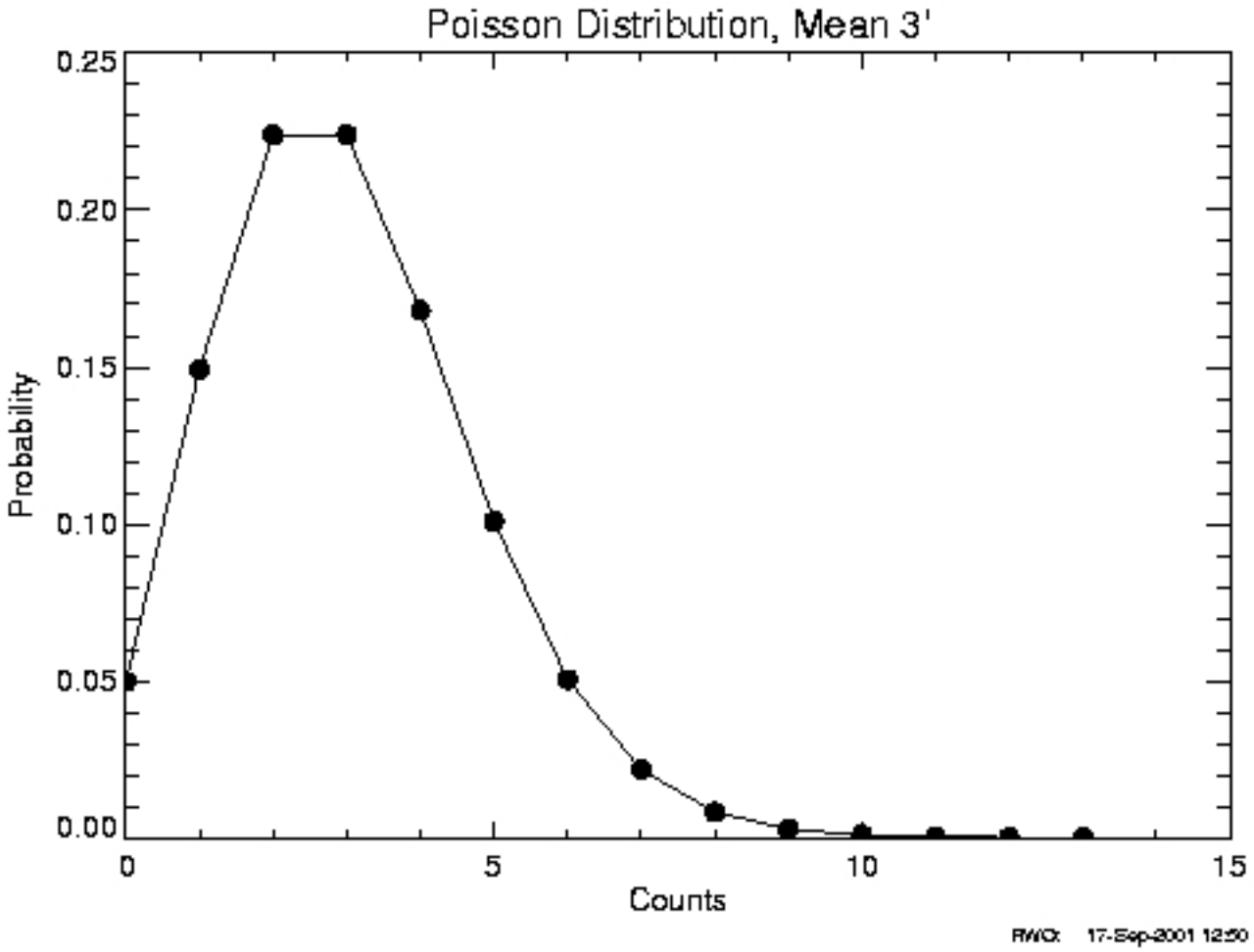
**Standard deviation: $\sigma = \sqrt{\mu}$**

**Implies mean/width $= \mu/\sqrt{\mu} = \sqrt{\mu}$**

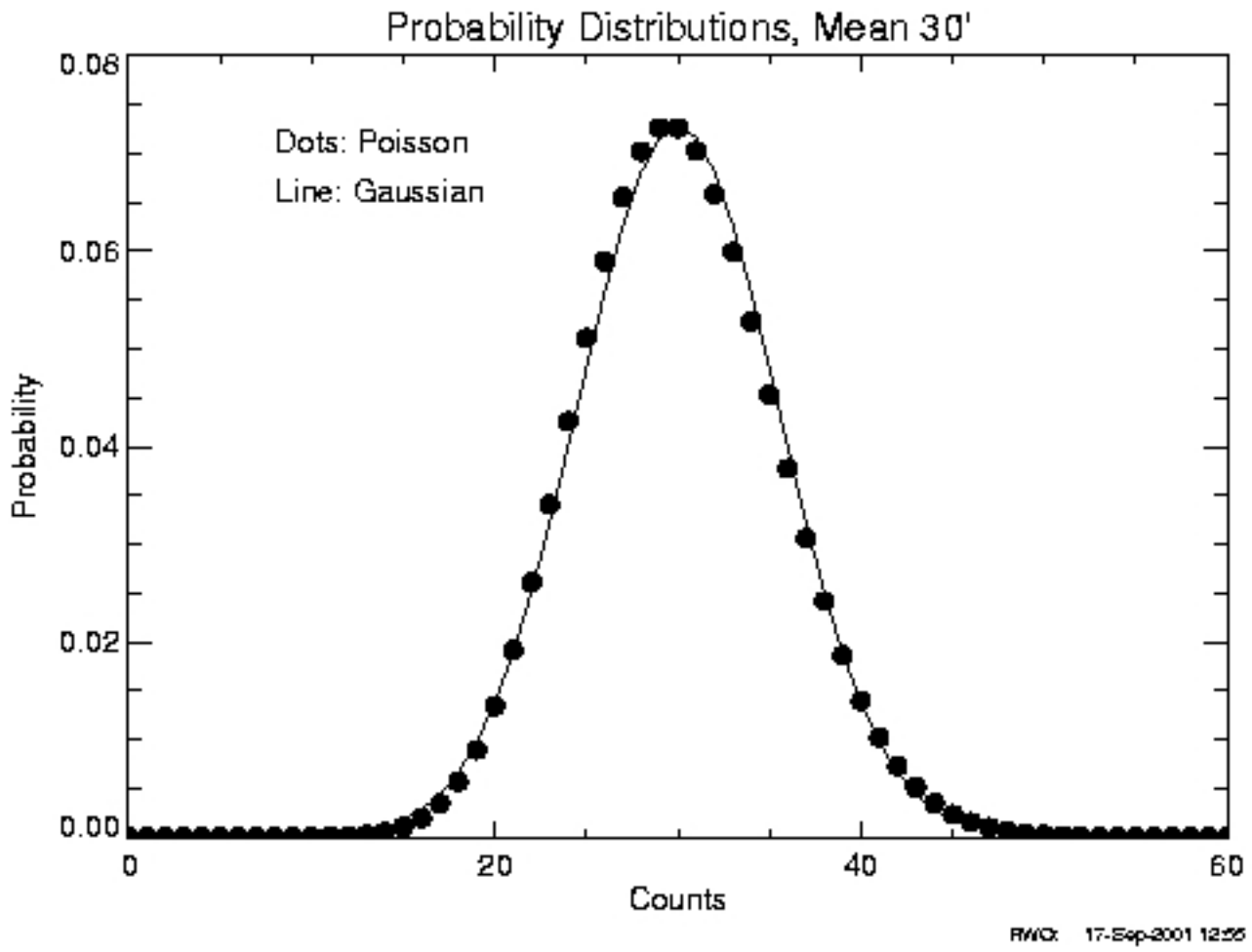
**$\rightarrow$ "Square root of n statistics"**

> **NB: the Poisson distribution is the proper description of a uniform counting process for small numbers of counts. For larger numbers ($n \gtrsim 30$), the Gaussian distribution is a good description and is easier to compute.**

# SAMPLE POISSON DISTRIBUTION



# POISSON AND GAUSSIAN COMPARED



# Gaussian Probability Distribution

**The Gaussian, or "normal," distribution is the limiting form of the Poisson distribution for large $\mu$ ($\gtrsim 30$)**

**Probability distribution:**

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

**Properties:**

$$\int_{-\infty}^{+\infty} p_G(x)\,dx = 1$$

**"Bell-shaped" curve; symmetric about mode at $\mu$**

**Mean value: $\mu$ (= median and mode)**

**Variance: $\sigma^2$**

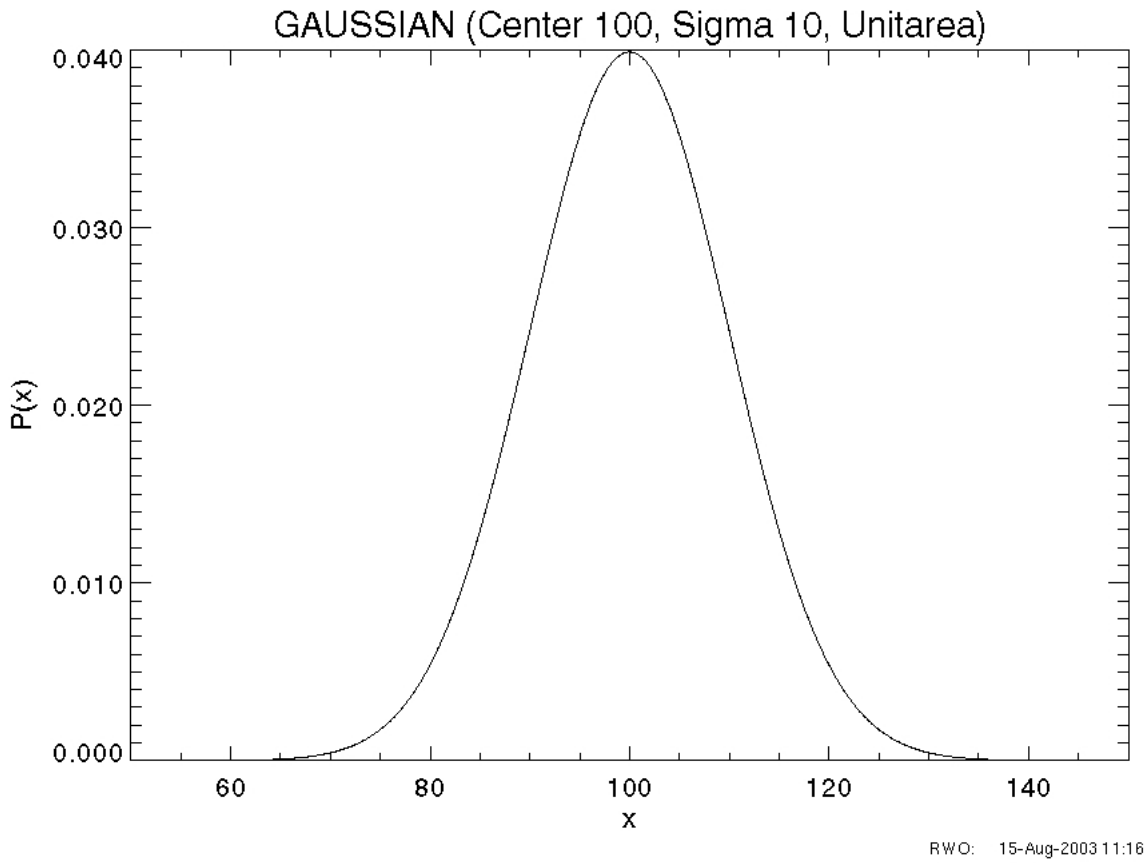
**Full Width Half Maximum $= 2.355\,\sigma$**

**If refers to a counting process ($x = n$ in bin), then $\sigma = \sqrt{\mu}$**

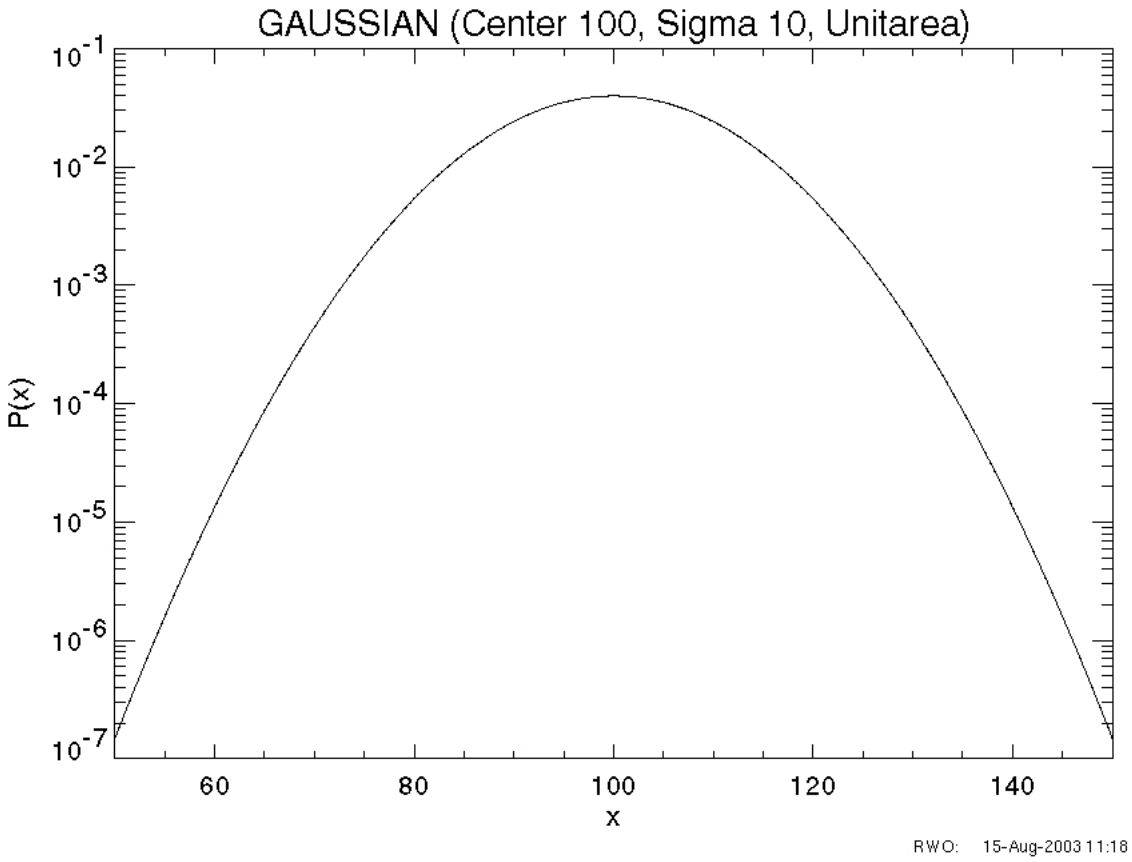
**Importance:**

**The "central limit theorem" of Gauss demonstrates that a Gaussian distribution applies to any situation where a large number of independent random processes contribute to the result. This means it is a valid statistical description of an enormous range of real-life situations. Much of the statistical analysis of data measurement is based on the assumption of Gaussian distributions.**

# SAMPLE GAUSSIAN DISTRIBUTION (Linear)



# SAMPLE GAUSSIAN DISTRIBUTION (Log)



# Chi-Square Probability Distribution

**The Chi-Square ($\chi^2$) function gives the probability distribution for any quantity which is the sum of the squares of independent, normally-distributed variables with unit variance. In the method of maximum likelihood it is important in testing the functional relationship between measured quantities.**

## Probability distribution:

$$p_\chi(\chi^2, \nu) = \frac{1}{2^{\nu/2}\,\Gamma(\nu/2)}\,(\chi^2)^{0.5(\nu-2)}\,\exp[-\chi^2/2]$$

**...where the Gamma function is defined as follows:**

$\Gamma(n+1) = n!$ **if** $n$ **is an integer**

$\Gamma(1/2) = \sqrt{\pi}$ **and** $\Gamma(n+1) = n\Gamma(n)$ **if** $n$ **is half-integer**

## Properties:

**Only one parameter, $\nu$, the "number of degrees of freedom." $\nu =$ the number of independent quantities in the sum of squares.**

**Mean and mode: $\nu$. Variance: $2\nu$**

**Asymmetric distribution**

# Sampling Theory

**In practice, we usually do not know the parameters of the parent distribution because this requires a very large number of measures. Instead, we try to make inferences about the parent distribution from finite (& often small) samples. Sampling theory describes how to estimate the moments of $p(x)$.**

**The results here are based on applying the "method of maximum likelihood" to variables whose parent distribution is assumed to be stationary and normally distributed.**

**Suppose we obtain a sample consisting of $M$ measurements of a given variable characterized by a normal distribution (with mean $\mu$ and standard deviation $\sigma$). Define the following two estimators:**

- **Sample mean:** $\bar{x} \equiv \frac{1}{M} \sum_{i=1}^{M} x_i$
- **Sample variance:** $s^2 \equiv \frac{1}{M-1} \sum_{i=1}^{M} (x_i - \bar{x})^2$

**These two estimators have the property that as $M \to \infty$,**

$$\bar{x} \to \mu \text{ and } s^2 \to \sigma^2$$

## SAMPLING THEORY (cont)

**How well determined is $\bar{x}$?**

**The "uncertainty" in $\bar{x}$ is its variance. But this is <u>not</u> the same as the variance in $x$. $\bar{x}$ is a random variable, and its variance can be computed as follows:**

$$s_{\bar{x}}^2 \equiv Var(\bar{x}) = \frac{1}{M^2}\sum_{i=1}^{M} Var(x_i) = \frac{1}{M}Var(x)$$

$$s_{\bar{x}}^2 \sim \frac{1}{M}s^2$$

$$s_{\bar{x}}^2 \sim \frac{1}{M(M-1)}\sum_{i=1}^{M}(x_i - \bar{x})^2$$

**$s_{\bar{x}}$ is known as the "standard error of the mean"**

**Important! $s_{\bar{x}} << \sigma$ if M is large.**

**The distinction between $\sigma$ and $s_{\bar{x}}$ is often overlooked by students and can lead to flagrant overestimation of errors in mean values.**

**The mean of a random variable can be determined very precisely regardless of its variance. This demonstrates the importance of repeated measurements...if feasible.**

# SAMPLING THEORY (cont)

## Probability Distribution of $\bar{x}$:

**By the central limit theorem, if we repeat a set of $M$ measures from a given parent distribution a large number of times, the resulting distribution of $\bar{x}_M$ will be a normal distribution regardless of the form of the parent distribution $p(x)$. It will have a standard deviation of $\sigma/\sqrt{M}$.**

## Inhomogeneous samples:

**A sample is inhomogeneous if $\sigma$ of the parent distribution is different for different measurements. This could happen with a long series of photometric determinations of a source's brightness, for instance.**

**Here, the values entering the estimates of the sample mean and variance must be weighted in inverse proportion to their uncertainties. The following expressions assume that the variance of each measurement can be estimated in some independent way:**

**Sample mean:** $$\bar{x} = \sum_{i=1}^{M} w_i x_i \Big/ \sum_{i=1}^{M} w_i$$

**Variance of the mean:** $$s_{\bar{x}}^2 = 1 \Big/ \sum_{i=1}^{M} w_i$$

**... where** $w_i = \frac{1}{\sigma_i^2}$

# The "Signal-to-Noise Ratio" (SNR) for Flux Measurements

**We adopt the sample mean $\bar{x}$ as the best estimate of the flux and $s_{\bar{x}}$, the standard error of the mean (not the standard deviation of the parent distribution), as the best estimate of the uncertainty in the mean flux.**

**Our working definition of signal-to-noise ratio is then:**

$$SNR \equiv \bar{x}/s_{\bar{x}}$$

**$s_{\bar{x}}$ here must include all effects which contribute to random error in the quantity $x$.**

**This is a basic "figure of merit" that should be considered in both planning observations (based on expected performance of equipment) and in evaluating them after they are made.**

# Propagation of Variance to Functions of Measured Variables

**If $u = f(x, y)$ is a function of two random variables, $x$ and $y$, then we can propagate the uncertainty in $x$ and $y$ to $u$ as follows:**

$$\sigma_u^2 = \sigma_x^2 \left(\frac{\partial u}{\partial x}\right)^2 + \sigma_y^2 \left(\frac{\partial u}{\partial y}\right)^2 + 2\sigma_{xy}^2 \left(\frac{\partial u}{\partial x}\right)\left(\frac{\partial u}{\partial y}\right)$$

**where the "covariance" of $x$ and $y$ is defined as**

$$\sigma_{xy} \equiv \lim_{M\to\infty} \frac{1}{M} \sum_i [(x_i - \bar{x})(y_i - \bar{y})]$$

**For <u>independent</u> random variables, $\sigma_{xy} = 0$.**

**So, we obtain for the following simple functions:**

$Var(kx) = k^2\, Var(x)$ **if $k$ is a constant**

$Var(x + y) = Var(x) + Var(y) + 2\sigma_{xy}^2$

$Var(xy) = y^2\, Var(x) + x^2\, Var(y) + 2xy\sigma_{xy}^2$

# Confidence Intervals

**A "confidence interval" is a range of values which can be expected to contain a given parameter (e.g. the mean) of the parent distribution with a specified probability. The smaller the confidence interval, the higher the precision of the measure.**

**(A) In the ideal case of a single measurement drawn from a normally-distributed parent distribution of known mean and variance, confidence intervals for the mean in units of $\sigma$ are easy to compute in the following form:**

$$P(\pm k\sigma) = \int_{\mu-k\sigma}^{\mu+k\sigma} p_G(x, \mu, \sigma)dx$$

**where $p_G$ is the Gaussian distribution. Results from this calculation are as follows:**

| $k$ | $P(\pm k\sigma)$ |
|---|---|
| 0.675 | 0.500 |
| 1.0 | 0.683 |
| 2.0 | 0.954 |
| 3.0 | 0.997 |

**Intepretation: A single measure drawn from this distribution will fall within $1.0\,\sigma$ of the mean value in 68% of the samples. Only 0.3% of the samples would fall more than $3.0\,\sigma$ from the mean.**

## CONFIDENCE INTERVALS (cont)

**(B) In the real world, we have only estimates of the properties of the parent distribution based on a finite sample. The larger the sample, the better the estimates, and the smaller the confidence interval.**

**To place confidence intervals on the estimate of the parent mean ($\mu$) based on a finite sample of $M$ measures, we use the probability distribution of the "Student" $t$ variable:**

$$t = (\bar{x} - \mu)\sqrt{M}/s$$

**where $s^2$ is the sample variance. The probability distribution of $t$ depends on the number of degrees of freedom, which in this case is $M - 1$. The probability that the true mean of the parent distribution lies within $\pm t\, s_{\bar{x}}$ of the sample mean is estimated by integrating the Student t-distribution from $-t$ to $+t$.**

| | $P(\mu \in \bar{x} \pm t\, s_{\bar{x}})$ | | |
|---|---|---|---|
| $t$ | $M = 2$ | $M = 10$ | $M = \infty$ |
| 0.5 | 0.295 | 0.371 | 0.383 |
| 0.6745 | 0.377 | 0.483 | 0.500 |
| 1.0 | 0.500 | 0.657 | 0.683 |
| 2.0 | 0.705 | 0.923 | 0.954 |
| 3.0 | 0.795 | 0.985 | 0.997 |

## CONFIDENCE INTERVALS (cont)

**Interpretation & comments on the $t$-distribution results:**

- **Entries for $M = \infty$ correspond to those for the Gaussian parent distribution quoted earlier, as expected.**
- **Values for small $M$ can be very different than for $M = \infty$. The number of observations is an important determinant of the quality of the measurement.**
- **The entry for 0.6745 is included because the formal definition of the "probable error" is $0.6745\, s_{\bar{x}}$. For a large number of measures, the probable error defines a 50% confidence interval. But for small samples, it is a very weak constraint.**
- **A better measure of uncertainty is the standard error of the mean, $s_{\bar{x}}$, which provides at least a 50% confidence interval for all $M$.**
- **Careful authors often quote "$3\,\sigma$" confidence intervals. This corresponds to $t = 3$ and provides 80% confidence for two measures and 99.7% for many measures. It is a strong contraint on results of a measurement.**

- **NB; the integrals in the preceding table were derived from an IDL built-in routine. The table contains output from the IDL statement:**
  `P = 2*T_PDF(t,M-1)-1.`

# GOODNESS OF FIT ($\chi^2$ TEST)

**Widely used standard for comparing an observed distribution with a hypothetical functional relationship for two or more related random variables. Determines the likelihood that the observed deviations between the observations and the expected relationship occur by chance. Assumes that the measuring process is governed by Gaussian statistics.**

**Two random variables $x$ and $y$. Let $y$ be a function of $x$ and a number $k$ of additional parameters, $\alpha_j$: $y = f(x; \alpha_1 ... \alpha_k)$.**

1. **Make $M$ observations of $x$ and $y$.**
2. **For each observation, estimate the total variance in the $y_i$ value, $\sigma_i^2$**
3. **We require $f(x; \alpha_1 ... \alpha_k)$. Either this must be known a priori, or it must be estimated from the data (e.g. by least squares fitting).**
4. **Then define**

$$\chi_0^2 = \sum_i^M \left( \frac{y_i - f(x_i)}{\sigma_i} \right)^2$$

5. **The probability distribution for $\chi^2$ was given earlier. It depends on the number of degrees of freedom $\nu$. If the $k$ parameters were estimated from the data, then $\nu = M - k$.**
6. **The predicted mean value of $\chi^2$ is $\nu$.**

7. **The integral $P_0 = \int_{\chi_0^2}^{\infty} p(\chi^2, \nu) d\chi^2$ then determines the probability that this or a higher value of $\chi_0^2$ would occur by chance.**
8. **The larger is $P_0$, the more likely it is that $f$ is correct. Values over 50% are regarded as consistent with the hypothesis that $y = f$.**
9. **Sample values of $\chi^2$ yielding a given $P_0$:**

| $P_0$ | $\nu = 1$ | $\nu = 10$ | $\nu = 200$ |
|---|---|---|---|
| 0.05 | 3.841 | 1.831 | 1.170 |
| 0.10 | 2.706 | 1.599 | 1.130 |
| 0.50 | 0.455 | 0.934 | 0.997 |
| 0.90 | 0.016 | 0.487 | 0.874 |
| 0.95 | 0.004 | 0.394 | 0.841 |

10. **Generally, one uses $1 - P_0$ as a criterion for <u>rejection</u> of the validity of $f$:**

    **E.g. if $P_0 = 5\%$, then with 95% confidence one can reject the hypothesis that $f$ is the correct description of $y$.**

11. **Important caveat: the $\chi^2$ test is <u>ambiguous</u> because it makes 2 assumptions: that $f$ is the correct description of $y(x)$ <u>and</u> that a Gaussian process with the adopted $\sigma$'s properly described the measurements. It will reject the hypothesis if <u>either</u> condition fails.**