2.830J / 6.780J / ESD.63J Control of Manufacturing Processes (SMA 6303)
Spring 2008

Control of Manufacturing Processes

Subject 2.830/6.780/ESD.63

Spring 2008

Lecture #5

Probability Models, Parameter Estimation, and Sampling

February 21, 2008

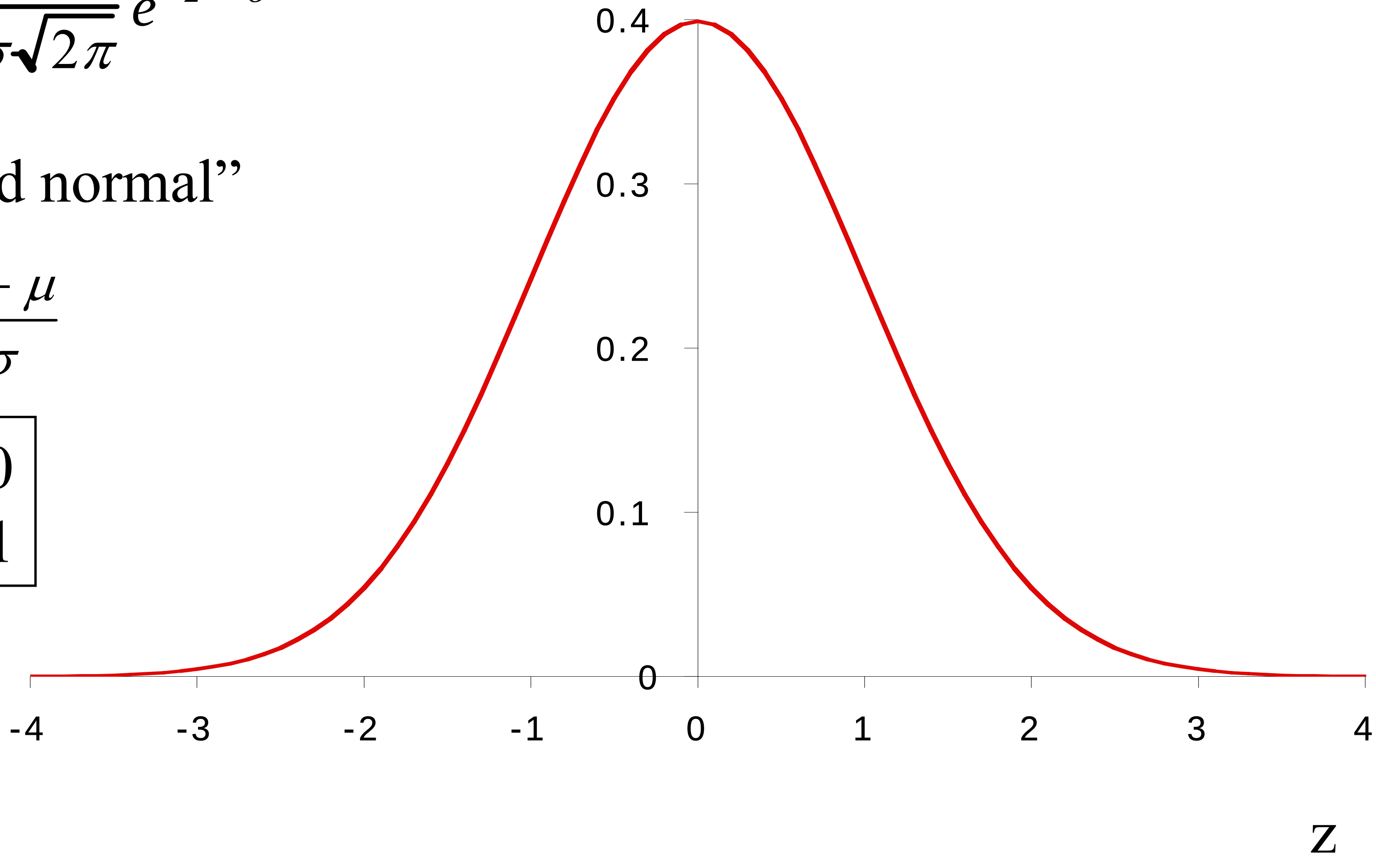
The Normal Distribution

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

"Standard normal"

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

$$\boxed{\begin{array}{l}\mu_z = 0 \\ \sigma_z = 1\end{array}}$$



Properties of the Normal pdf

- Symmetric about mean
- Only two parameters:

  μ and σ

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

- Mean (μ) and Variance (σ^2) have well known "estimators" (average and sample variance)

Testing the Model: e.g. Is the Process "Normal" ?

- Is the underlying distribution really normal?
  - Look at histogram
  - Look at curve fit to histogram
  - Look at % of data in 1, 2 and 3σ bands
    - Confidence Intervals
  - Look at "kurtosis"
    - Measure of deviation from normal
  - Probability (or qq) plots (see Mont. 3-3.7 or MATLAB stats toolbox)

Kurtosis: Deviation from Normal

$$k = \frac{E(x - \mu_x)^4}{\sigma^4} - 3$$

$k = 0$ - normal

$k > 0$ - more "peaked"

$k < 0$ - more "flat"

For sampled data:

$$k = \left[\frac{n(n+1)}{(n-1)(n-2)(n-3)} \sum \left(\frac{x_i - \bar{x}}{s} \right)^4 \right] - \frac{3(n-1)^2}{(n-2)(n-3)}$$

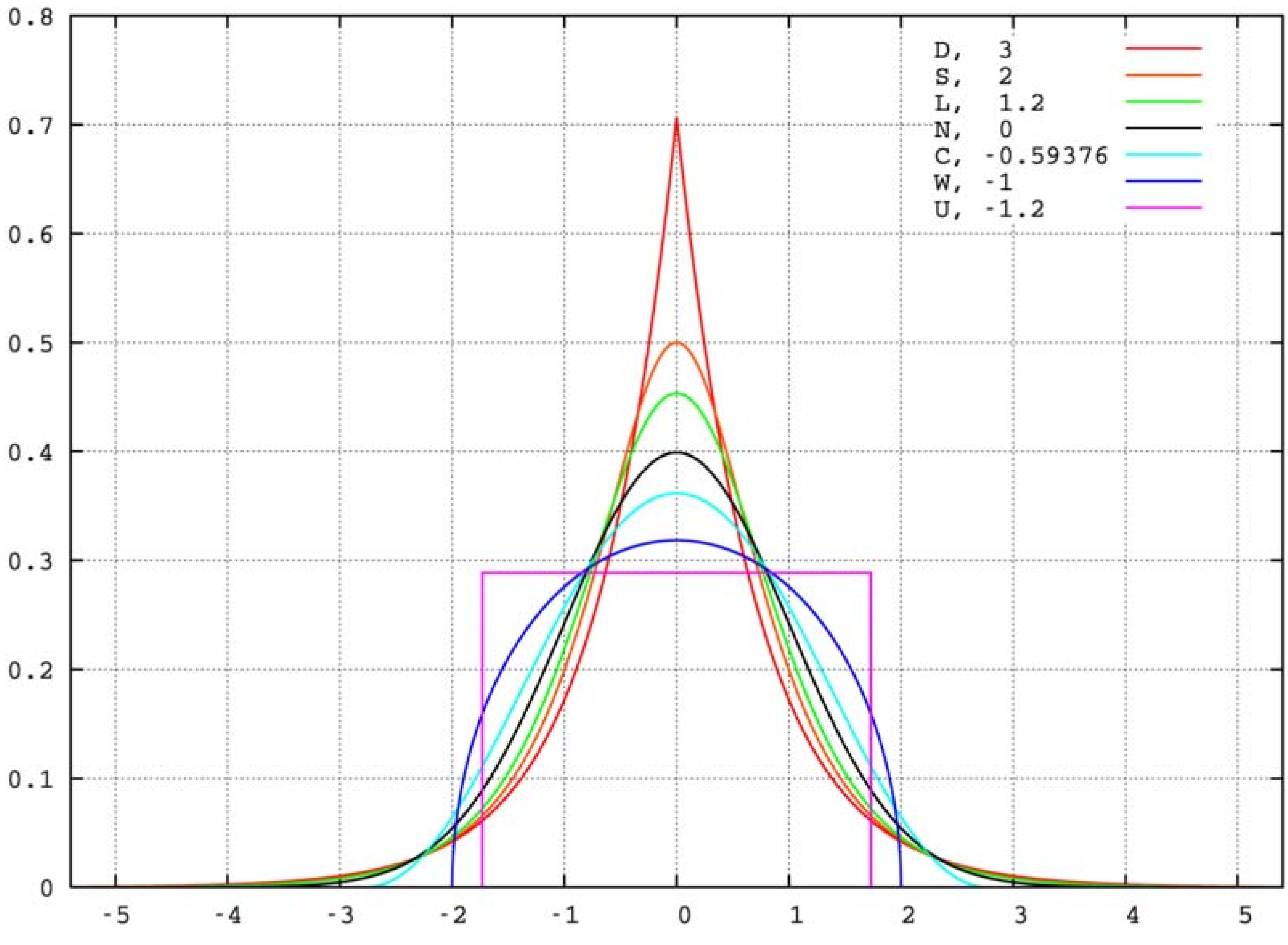
Kurtosis for Some Common Distributions

D: Laplace (k = 3)

L: logistic (k = 1.2)

N: normal (k = 0)

U: uniform (k = -1.2)



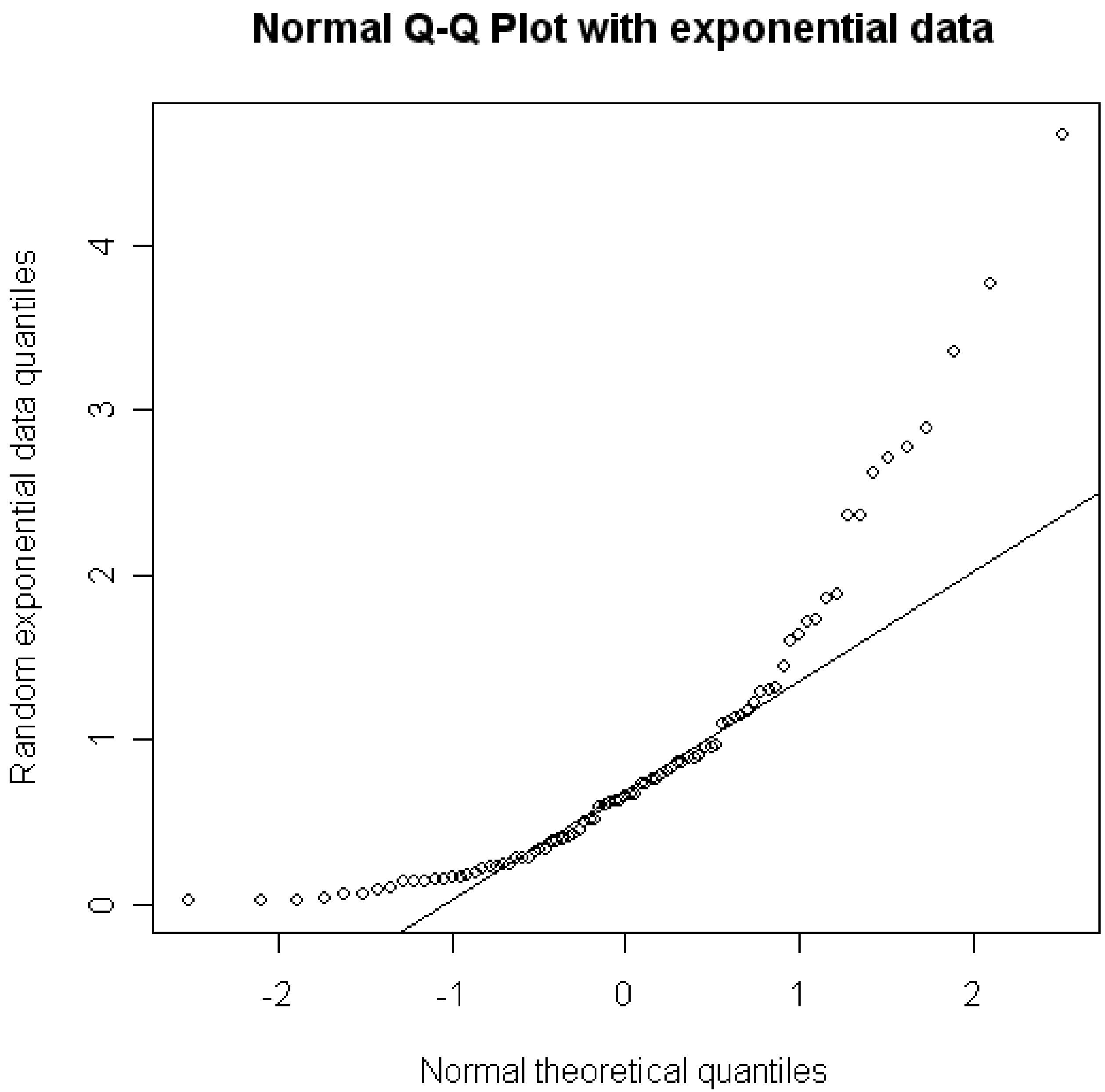
Source: Wikimedia Commons, http://commons.wikimedia.org

Manufacturing

Quantile-Quantile (qq) Plots

- Plot
  - normalized (mean centered and scaled to *s*)

  vs.
  - theoretical position of unit normal distribution for ordered data
- Normal distribution: data should fall along line



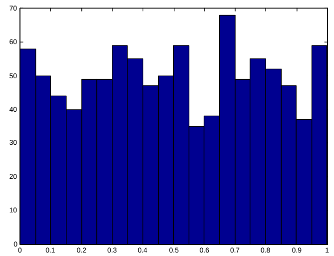
Source: Wikimedia Commons, http://commons.wikimedia.org

Guaranteeing "Normality"
The Central Limit Theorem

– If $x_1, x_2, x_3 \ldots x_N \ldots$ are N <u>independent</u> observations of a random variable with "moments" μ_x and σ^2_x,

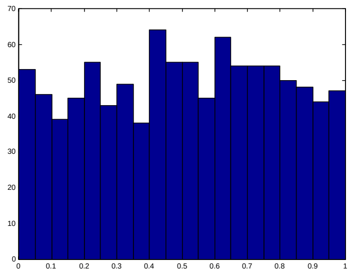
– The distribution of the ***sum*** of all the samples will tend toward normal.

Example: Uniformly Distributed Data



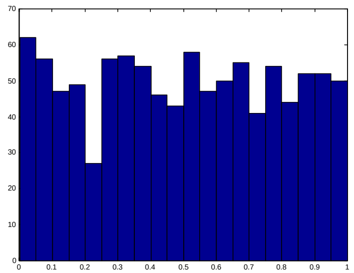
x_1

$+$



x_2

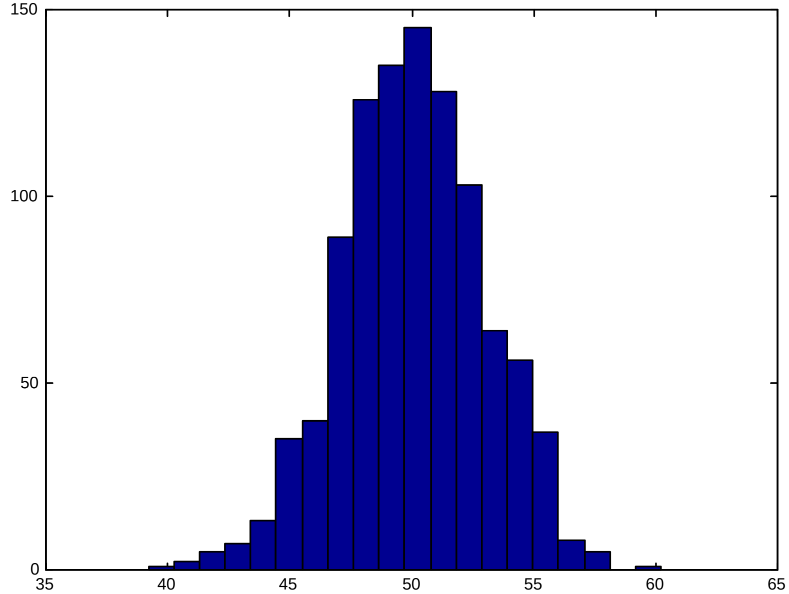
$+ \ldots$



x_{100}

Sum of 100 sets of 1000 points each

$$y = \sum_{i=1}^{100} x_i$$



Sampling: Using Measurements (Data) to Model the Random Process

- In general *p(x)* is unknown
- Data can suggest form of *p(x)*
  - e.g.. uniform, normal, weibull, etc.
- Data can be used to estimate parameters of distributions
  - e.g. μ and σ for normal distribution: $p(x) = N(\mu, \sigma^2)$

- How to estimate
  - Sample Statistics
- Uncertainty in estimates
  - Sample Statistic pdf's

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

Sample Statistics

$x_i = \quad n$ samples of x

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

Average or sample mean

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

Sample variance

$$s = \sqrt{s^2}$$

Sample standard deviation

Sample Mean Uncertainty

- If all x_i come from a distribution with μ_x and σ^2_x, *and we divide the sum by n:*

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

$$\bar{x} = c_1 x_1 + c_2 x_2 + c_3 x_3 + \ldots c_n x_n$$

$$c_i = \frac{1}{n}$$

Then: $\mu_{\bar{x}} = \mu_x$ and $\sigma^2_{\bar{x}} = \frac{1}{n}\sigma^2_x$ or $\boxed{\sigma_{\bar{x}} = \frac{1}{\sqrt{n}}\sigma_x}$

Manufacturing as Random Processes

- All physical processes have a degree of natural randomness
- We can model this behavior using probability distribution functions
- We can calibrate and evaluate the quality of this model from measurement data

Formal Use of Statistical Models

- **Discrete Variable Distributions and Uses**
  – Attribute Modeling
- **Sampling**: Key distributions arising in sampling
  - Chi-square, t, and F distributions
- **Estimation**:
  – Reasoning about the population based on a sample
- Some basic **confidence intervals**
  - Estimate of mean with variance known
  - Estimate of mean with variance not known
  - Estimate of variance
- **Hypothesis tests**

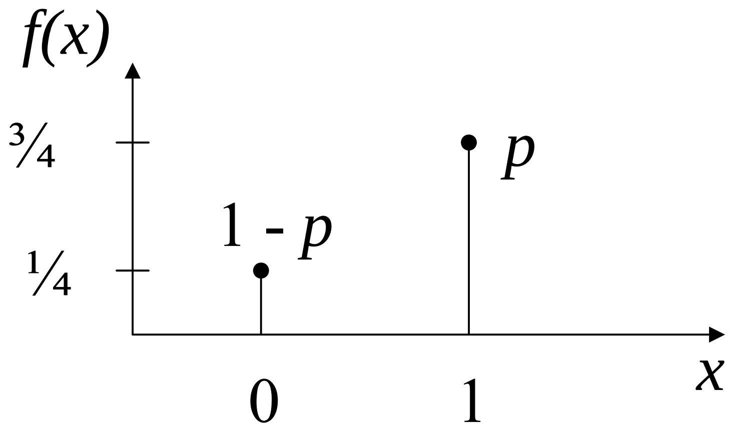
Discrete Distribution: Bernoulli

Bernoulli trial: an experiment with two outcomes

$$\begin{aligned} \Pr(\text{success}) &= \Pr(1) \\ \Pr(\text{failure}) &= \Pr(0) \end{aligned}$$

Probability density function (pdf):

$$f(x,p) = \begin{cases} p & x = 1 \\ 1-p & x = 0 \end{cases}$$



$$\mu = \mathrm{E}[f(x,p)] = 1 \cdot p + 0 \cdot (1-p) = p$$

$$\sigma^2 = \mathrm{Var}[f(x,p)] = p(1-p)$$

Discrete Distribution: Binomial

Repeated random Bernoulli trials

$$f(x,p,n) = \binom{n}{x} p^x (1-p)^{n-x}, \quad x = 0, 1, 2, ..., n$$

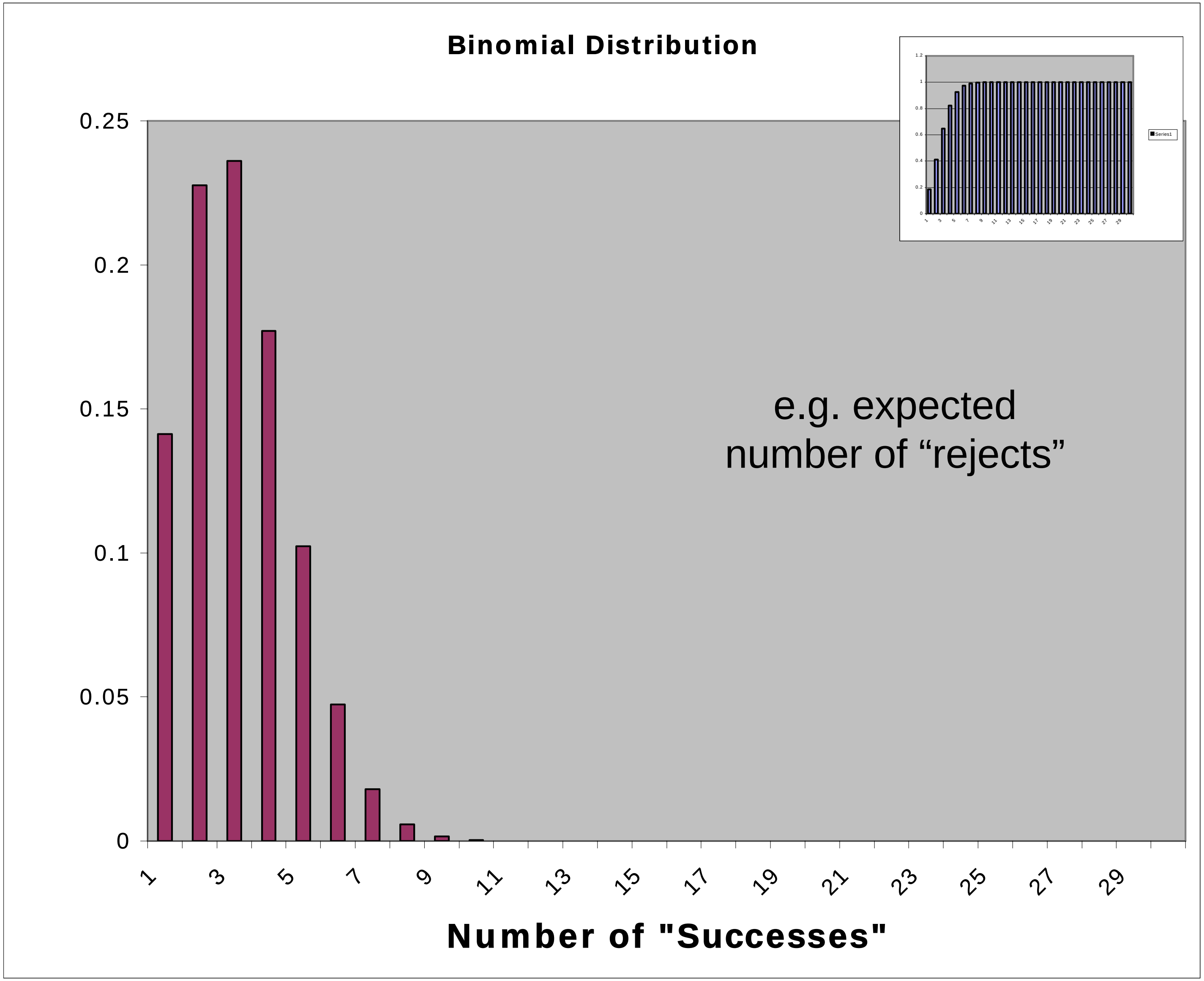
where $\binom{n}{x}$ is "n choose x" $= \frac{n!}{x!(n-x)!}$

$$\mu = np$$
$$\sigma^2 = np(1-p)$$

$x \sim B(n,p)$ where $\sim$ reads "is distributed as" a binomonial

- *n* is the number of trials
- *p* is the probability of "success" on any one trial
- *x* is the number of successes in n trials

Binomial Distribution



Discrete Distribution: Poisson

$$f(x, \lambda) = \frac{e^{-\lambda}\lambda^x}{\lambda!}, \quad \lambda = 0, 1, 2, \ldots \qquad\qquad x \sim P(\lambda)$$

Mean: $\mu = \lambda$

Variance: $\sigma^2 = \lambda$

Example applications:

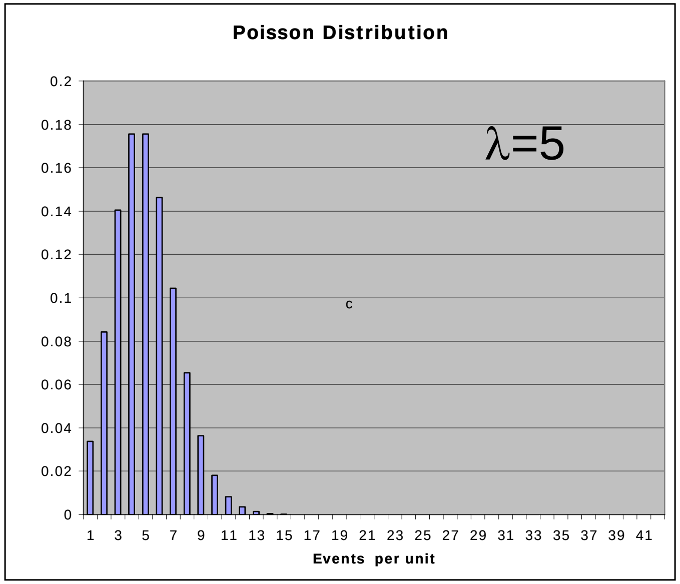
misprints on page(s) of a book

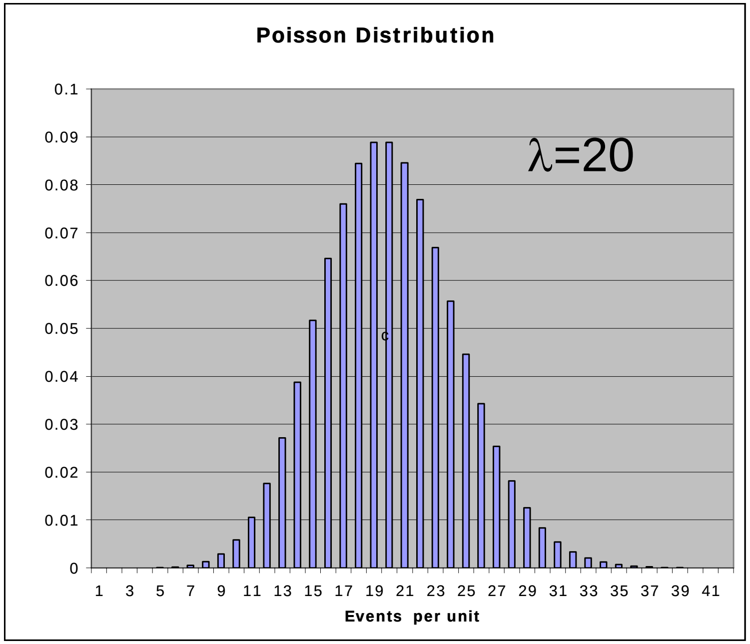
defects on a wafer

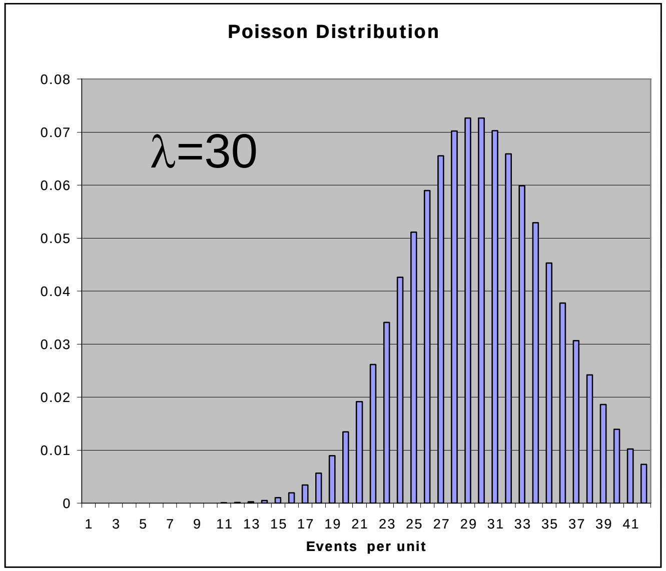
- Poisson is a good approximation to Binomial when *n* is large and *p* is small (< 0.1)

$$\mu = \lambda \approx np$$

Poisson Distributions







e.g. defects/device

Back to Continuous Distributions

- Uniform Distribution
- Normal Distribution
  - Unit (Standard) Normal Distribution

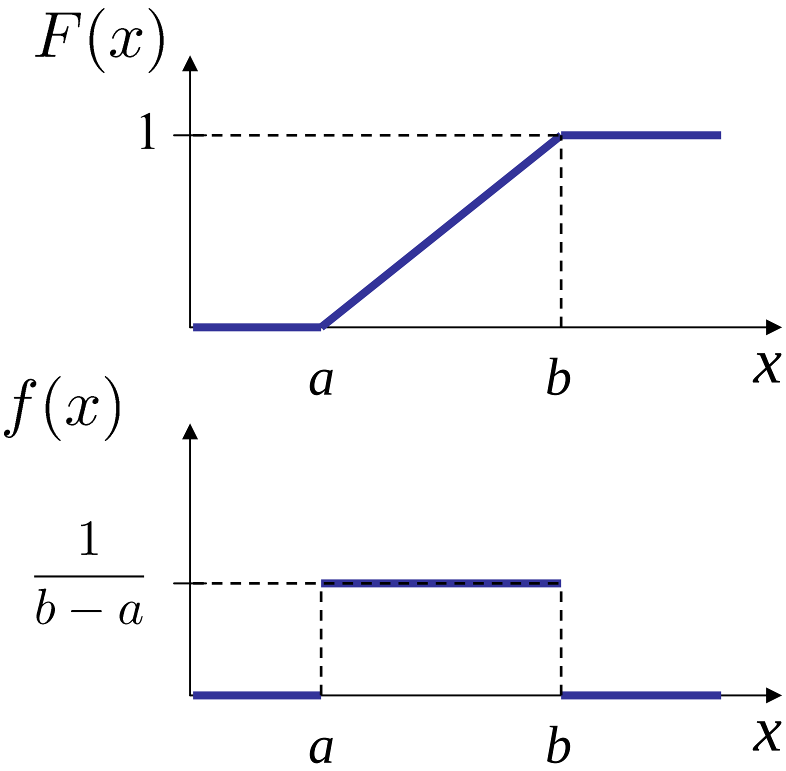
Continuous Distribution: Uniform

cdf

$$F(x) = \begin{cases} 0 & x < a \\ \frac{x-a}{b-a} & a \le x < b \\ 1 & x \ge b \end{cases}$$

pdf

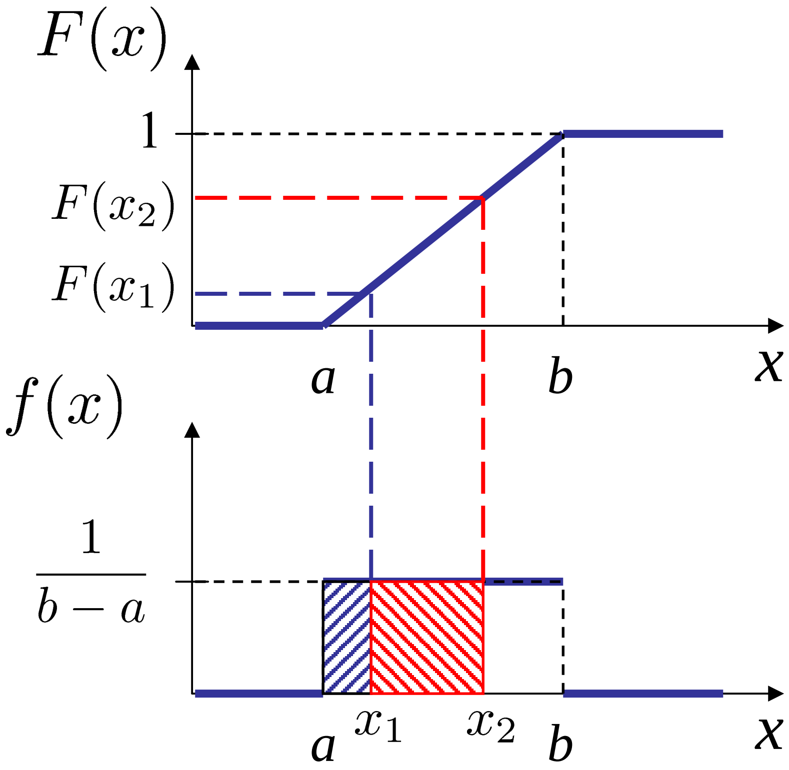
$$f(x) = \begin{cases} \frac{1}{b-a} & a \le x < b \\ o & \text{otherwise} \end{cases}$$



Standard Questions For a Known cdf or pdf

- Probability x less than or equal to some value

$$\Pr(x \le x_1) = \int_{-\infty}^{x_1} f(x)\ dx = F(x_1)$$

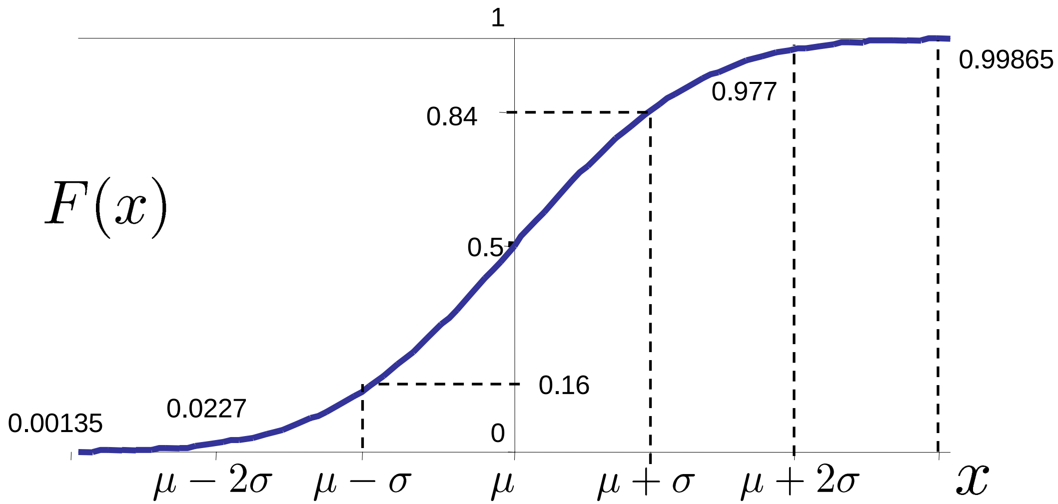


- Probability x sits within some range

$$\Pr(x_1 < x < x_2) = \int_{x_1}^{x_2} f(x)\ dx = F(x_2) - F(x_1)$$

Continuous Distribution: Normal or Gaussian

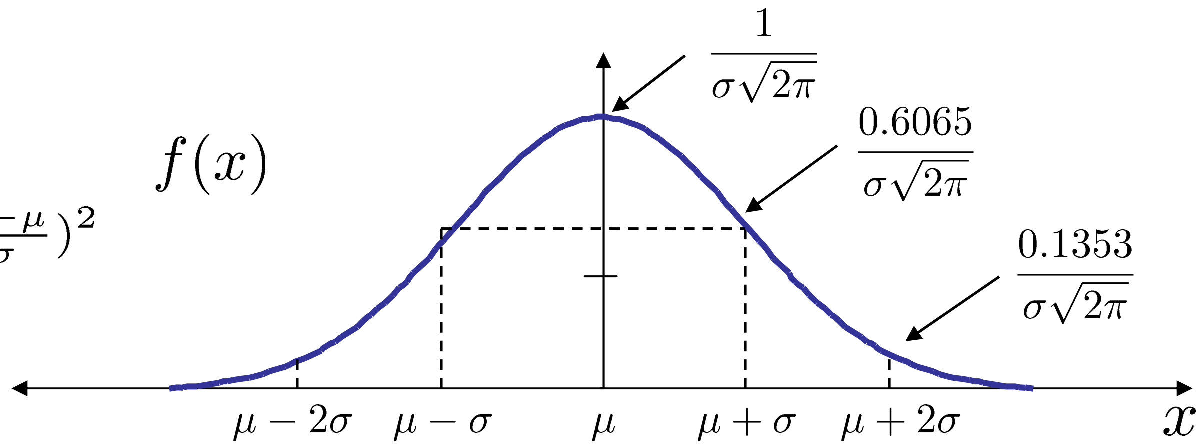
cdf



pdf

$$x \sim \mathrm{N}(\mu, \sigma^2)$$

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



Continuous Distribution: Unit Normal

- Normalization $$z = \frac{x-\mu}{\sigma} \qquad z \sim \mathrm{N}(0,1)$$

Mean $$\mathrm{E}(z) = 0$$

Variance $$\mathrm{Var}(z) = 1 \quad \Rightarrow \quad \mathrm{std.dev.}(z) = 1$$

pdf $$f(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2}$$

cdf $$F(z) = \Phi(z) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}v^2}\ dv$$

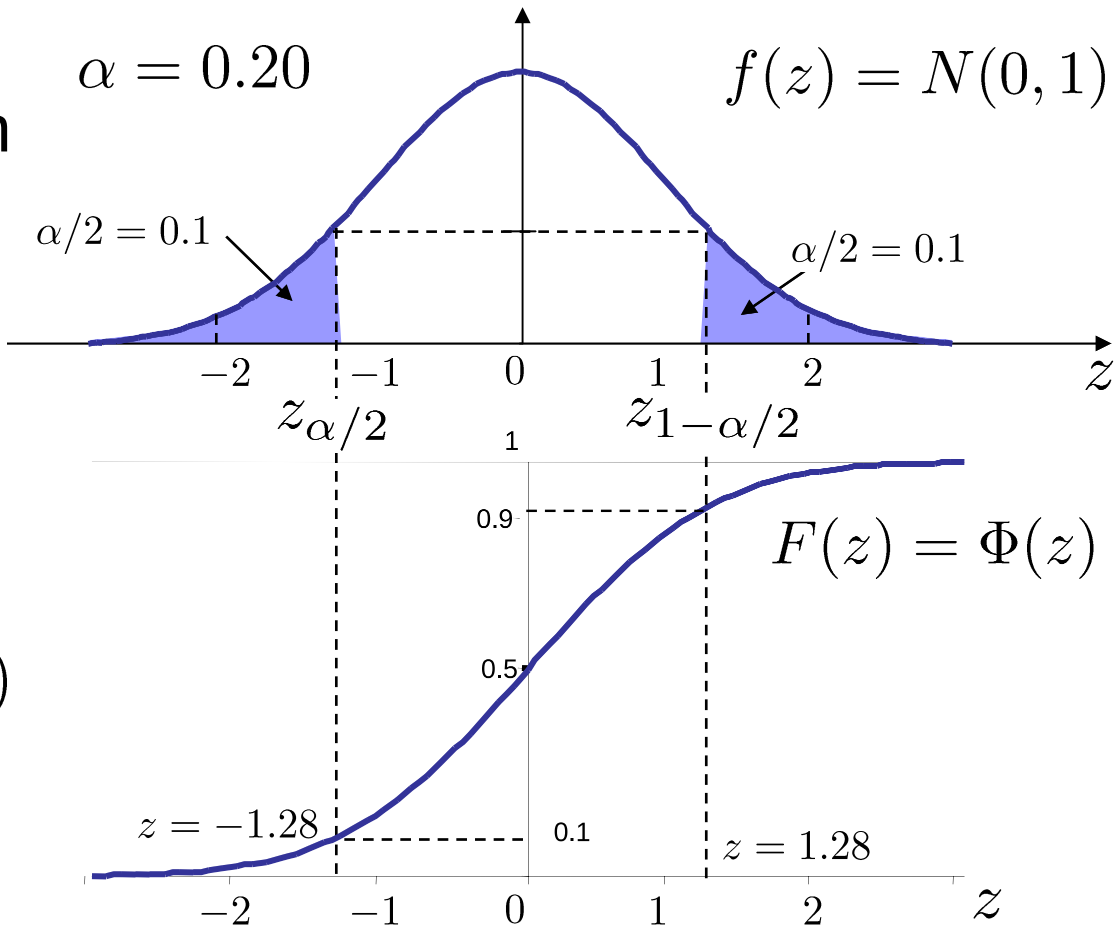
Using the Unit Normal pdf and cdf

- We often want to talk about “percentage points” of the distribution – portion in the tails

$$\Phi(z_{\alpha/2}) = \alpha/2$$
$$1 - \Phi(z_{\alpha/2}) = 1 - \alpha/2$$

$$z_{\alpha/2} = \Phi^{-1}(\alpha/2)$$
$$z_{1-\alpha/2} = -\Phi^{-1}(\alpha/2)$$

$$z_{0.10} = -1.28$$
$$z_{0.90} = 1.28$$



Use of the pdf: Location of Data

- How likely are certain values of the random variable?
- For a “Standard Normal” Distribution:

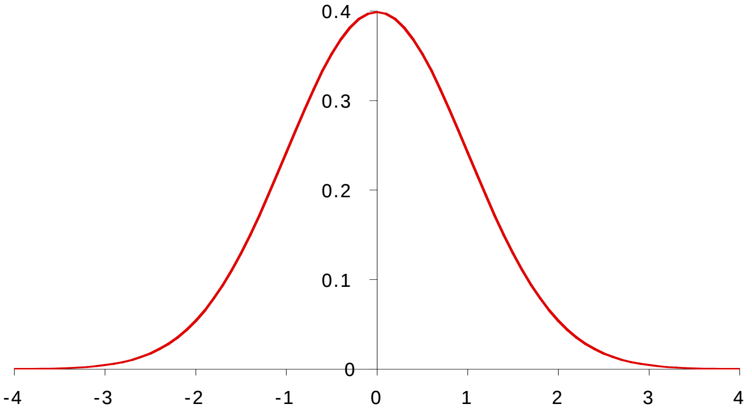
$$z = \frac{(x-\mu)}{\sigma}$$

N(0,1) $\mu=0$ $\sigma=1$

$$z = 1 \quad \Rightarrow \quad x = 1\sigma$$

$$z = 2 \quad \Rightarrow \quad x = 2\sigma$$

$$z = 3 \quad \Rightarrow \quad x = 3\sigma$$



Location of Data

$P(-1 \leq z \leq 1) = P(z \leq 1) - P(z \leq -1) = \Phi(1) - \Phi(-1)$

$(\pm 1\sigma)$ $= 0.841 - (1-0.841) = \mathbf{0.682}$

$P(-2 \leq z \leq 2) = P(z \leq 2) - P(z \leq -2) = 0.977 - (1-0.977) = \mathbf{0.954}$

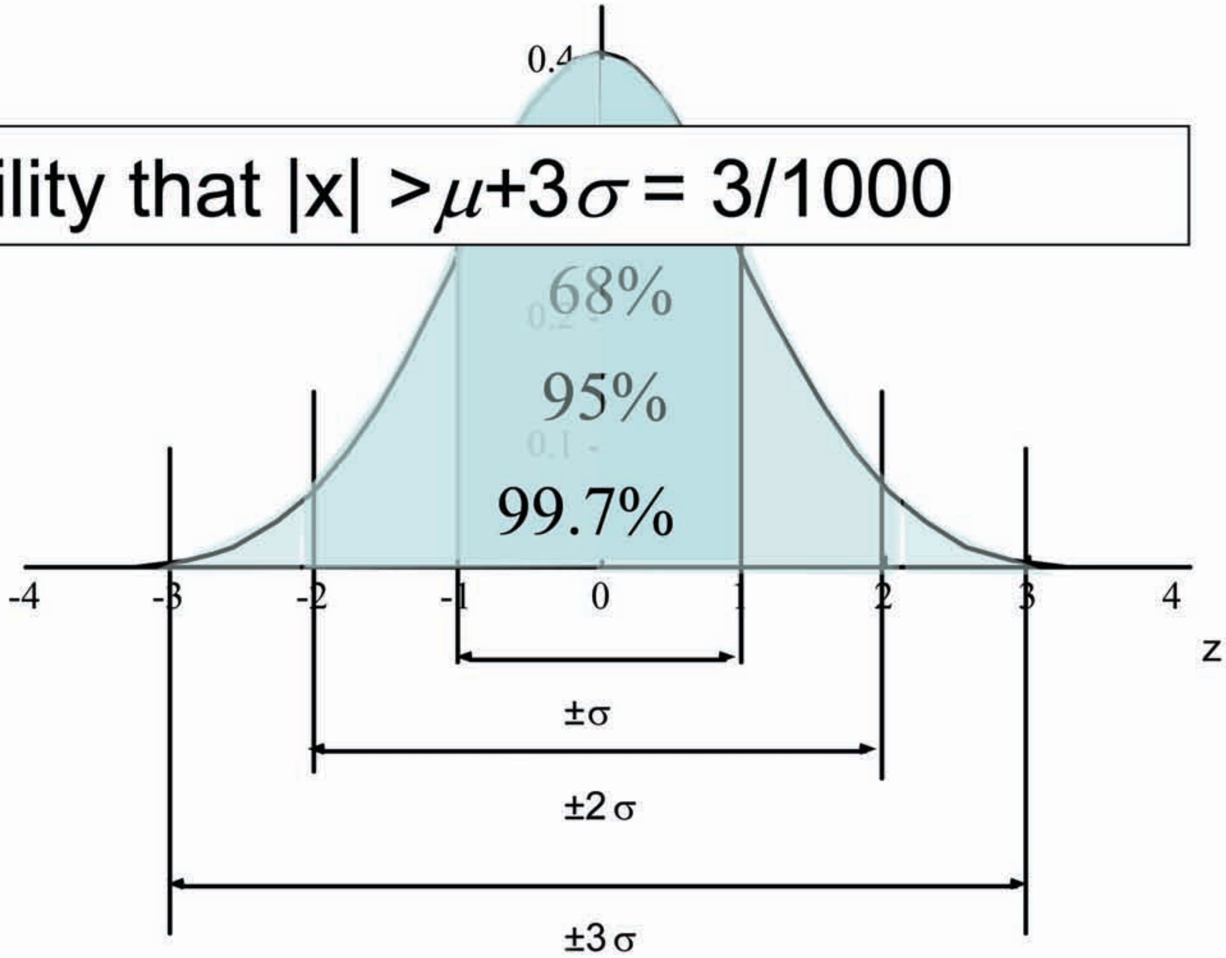
$(\pm 2\sigma)$

$P(-3 \leq z \leq 3) = P(z \leq 3) - P(z \leq -3) = 0.998 - (1-0.998) = \mathbf{0.997}$

$(\pm 3\sigma)$

$\Phi(z)$ tabulated (e.g. p. 752 of Montgomery)

Location of Data



- Probability that $|x| > \mu + 3\sigma = 3/1000$

Statistics

The field of statistics is about **reasoning** in the face of **uncertainty,** based on evidence from **observed data**

- Beliefs:
  - Probability distribution or probabilistic model form
  - Distribution/model parameters
- Evidence:
  - Finite set of observations or data drawn from a population (experimental measurements or observations)
- Models:
  - Seek to explain data wrt a model of their probability

Sampling to Determine Parameters of the Parent Probability Distribution

- Assume Process Under Study has a Parent Distribution *p(x)*
- Take "*n*" Samples From the Process Output (x_i)

- Look at Sample Statistics (e.g. sample mean and sample variance)
- Relationship to Parent
    - Both are Random Variables
    - Both Have Their Own Probability Distributions
- Inferences about the process (the parent distribution) via Inferences about the derived sampling distribution

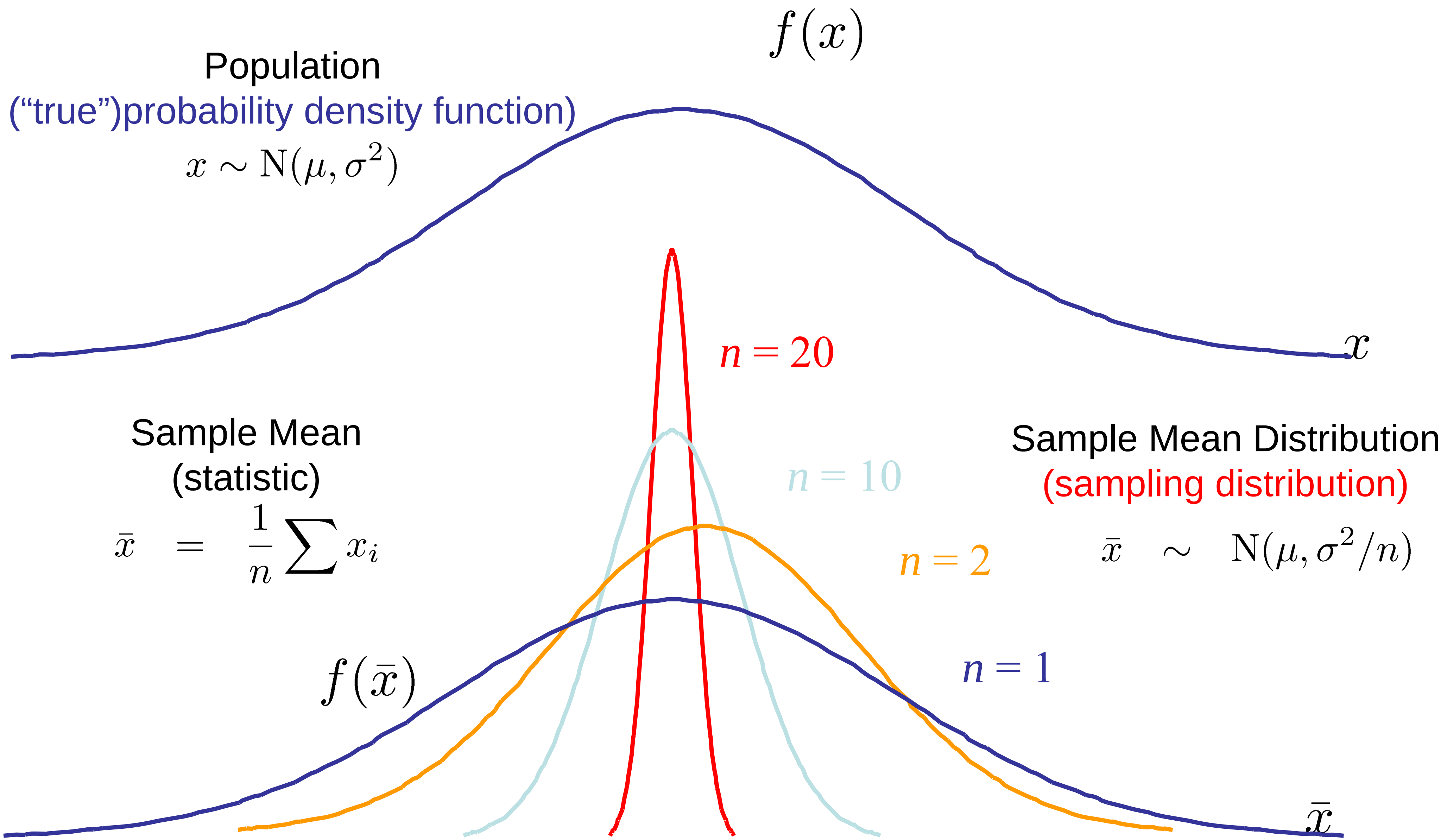
Moments of the Population vs. Sample Statistics

	Underlying model or Population Probability	Sample Statistics
• Mean	$\mu = \mu_x = \mathrm{E}(x)$	$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$
• Variance	$\sigma^2 = \sigma_{xx}^2 = \mathrm{E}[(x-\mu_x)^2]$	$s^2 = s_x^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$
• Standard Deviation	$\sigma = \sqrt{\sigma^2}$	$s = \sqrt{s^2}$
• Covariance	$\sigma_{xy}^2 = \mathrm{E}[(x-\mu_x)(y-\mu_y)] = \mathrm{E}(xy) - \mathrm{E}(x)\mathrm{E}(y)$	$s_{xy}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i = \bar{y})$
• Correlation Coefficient	$\rho_{xy} = \frac{\sigma^2 xy}{\sigma_x \sigma_y} = \frac{\mathrm{Cov}(xy)}{\sqrt{\mathrm{Var}(x)\mathrm{Var}(y)}}$	$r_{xy} = \frac{s^2 xy}{s_x s_y}$

Sampling and Estimation

- Sampling: act of making observations from populations
- Random sampling: when each observation is identically and independently distributed (IID)
- Statistic: a function of sample data; a value that can be computed from data (contains no unknowns)
  - Average, median, standard deviation
  - Statistics are by definition also random variables

Population vs. Sampling Distribution



Sampling and Estimation, cont.

- Sampling
- Random sampling
- Statistic
- A **statistic** is a random variable, which itself has a **sampling (probability) distribution**
  – I.e., if we take multiple random samples, the value for the statistic will be different for each set of samples, but will be governed by the same sampling distribution
- If we know the appropriate sampling distribution, we can **reason** about the underlying population based on the observed value of a statistic
  – E.g. we calculate a sample mean from a random sample; in what range do we think the actual (population) mean sits?

Sampling and Estimation – An Example

- Suppose we know that the thickness of a part is normally distributed with std. dev. of 10:

$$T \sim N(\mu_{\text{unknown}}, 100)$$

- We sample n = 50 random parts and compute the mean part thickness:

$$\bar{T} = \frac{1}{n}\sum_{i=1}^{n} T_i = 113.5$$

- First question: What is distribution of the mean of T = $\bar{T}$?

$$\bar{T} \sim N(\mu, 2)$$

$$\mathrm{E}(\bar{T}) = \mu$$
$$\mathrm{Var}(\bar{T}) = \sigma^2/n = 100/50$$
Normally distributed

- Second question: can we use knowledge of $\bar{T}$ distribution to reason about the actual (population) mean μ given observed (sample) mean?

Estimation and Confidence Intervals

- Point Estimation:
  - Find best values for parameters of a distribution
  - Should be
    - Unbiased: expected value of estimate should be true value
    - Minimum variance: should be estimator with smallest variance

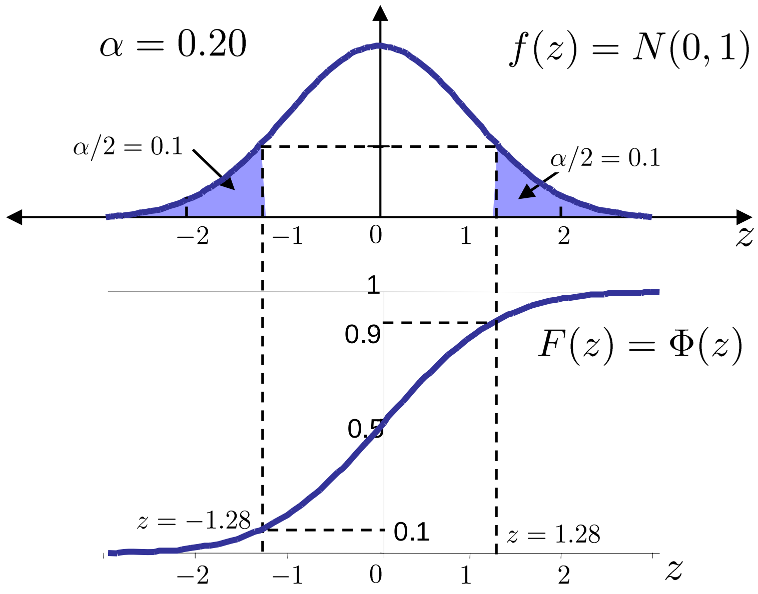
- Interval Estimation:
  - Give bounds that contain actual value with a given probability
  - Must know sampling distribution!

Confidence Intervals: Variance <u>Known</u>

- We know σ, e.g. from historical data
- Estimate mean in some interval to (1-α)100% confidence

$$\bar{x} - z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{x} + z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}$$

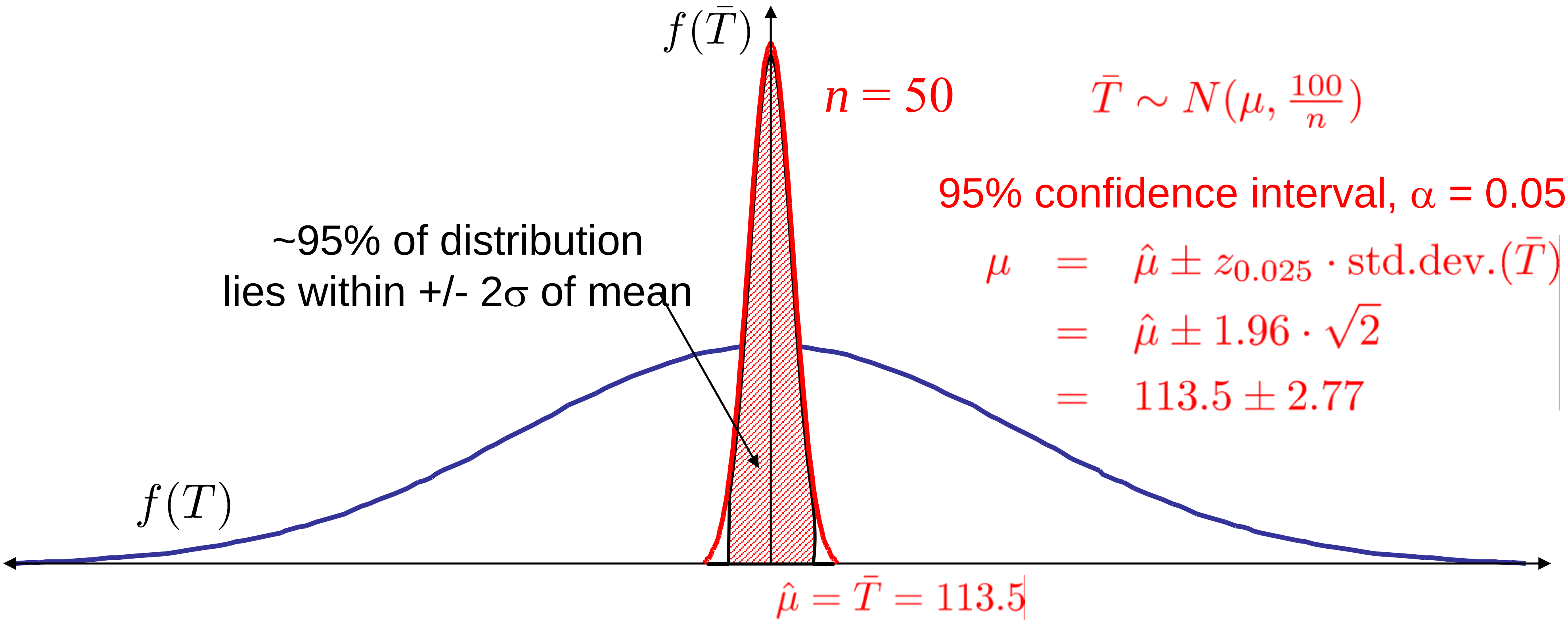
Remember the unit normal percentage points



Apply to the **sampling distribution** for the sample mean

Example, Cont'd

- Second question: can we use knowledge of $\bar{T}$ distribution to reason about the actual (population) mean μ given observed (sample) mean?



$\bar{T} \sim N(\mu, \frac{100}{n})$

95% confidence interval, α = 0.05

$$\begin{aligned} \mu &= \hat{\mu} \pm z_{0.025} \cdot \text{std.dev.}(\bar{T}) \\ &= \hat{\mu} \pm 1.96 \cdot \sqrt{2} \\ &= 113.5 \pm 2.77 \end{aligned}$$

Summary

- Process as Random Variable
  - Histograms to pdf's
- Different Distributions for Different Processes
  - Discrete or Binary (e.g. Defects)
  - Continuous (e.g. Dimensional Variation)
- Parent Distributions and Sampling
  - Estimating the Parent from Data
- Use of Distributions to establish "Confidence" on Parameter Estimates