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 #6

Sampling Distributions and Statistical Hypotheses

February 26, 2008

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/observations)
- Models:
  - Seek to explain data wrt a model of their probability

Topics

- Sampling Distributions (χ^2 and Student's-t)
  – Uncertainty of Parameter Estimates
  – Effect of Sample Size
  – Examples of Inference
- Inferences from Distributions
  – Statistical Hypothesis Testing
  – Confidence Intervals
- Hypothesis Testing
- The Shewhart Hypothesis and Basic SPC
  – Test statistics - xbar and S

Sampling to Determine 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 Process via Inferences about the Parent Distribution

Moments of the Population vs. Sample Statistics

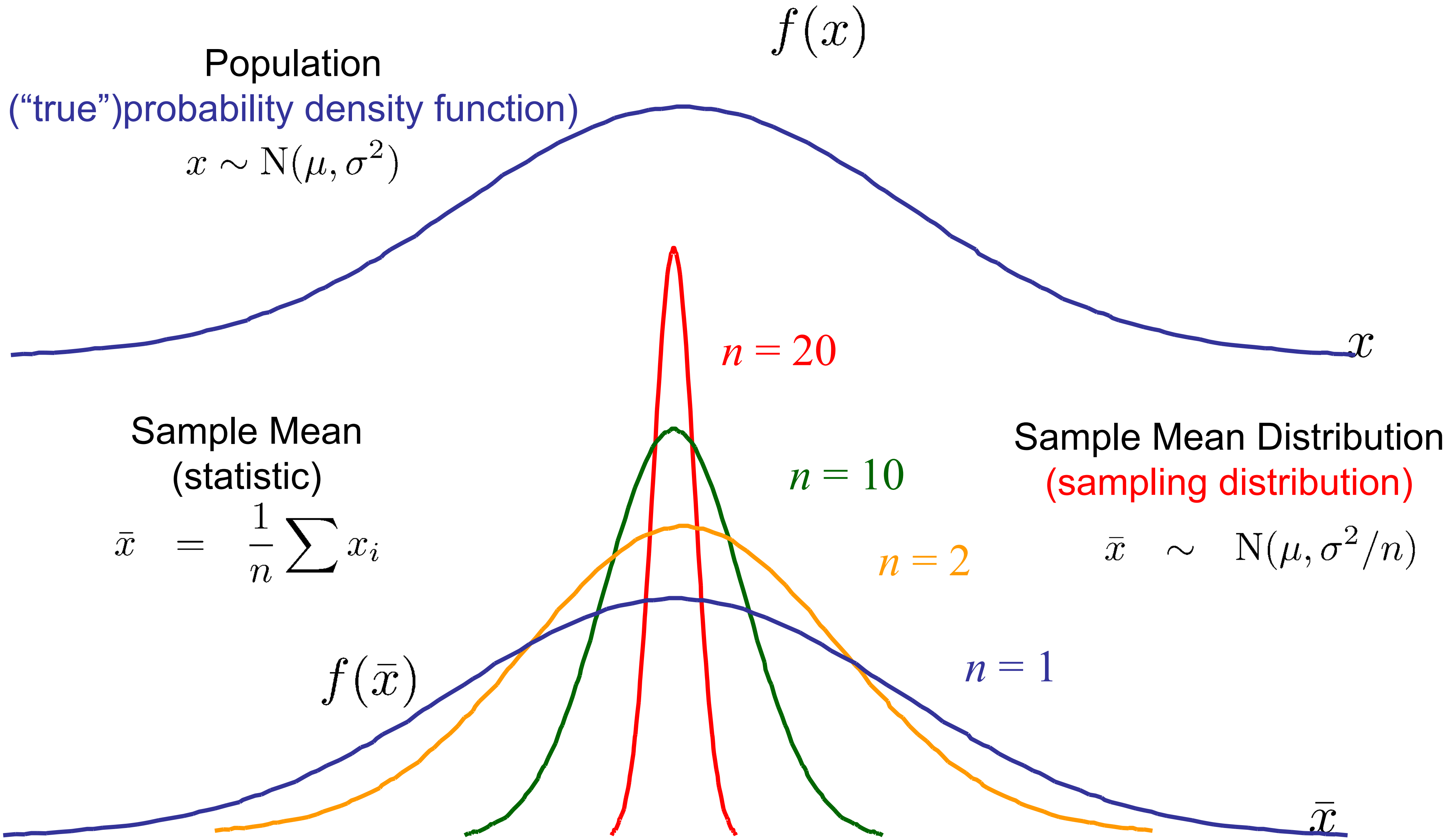
	Underlying model or Population Probability	Sample Statistics
• Mean	$\mu = \mu_x = \mathrm{E}(x)$	$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$
• Variance	$\sigma^2 = \sigma_{xx}^2 = \mathrm{E}[(x-\mu_x)^2]$	$s^2 = s_x^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$
• Standard Deviation	$\sigma = \sqrt{\sigma^2}$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

Manufacturing

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



Manufacturing

Sampling and Estimation, cont.

- 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 really sits?

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!

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 = $\bar{T}$?

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

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

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

Manufacturing

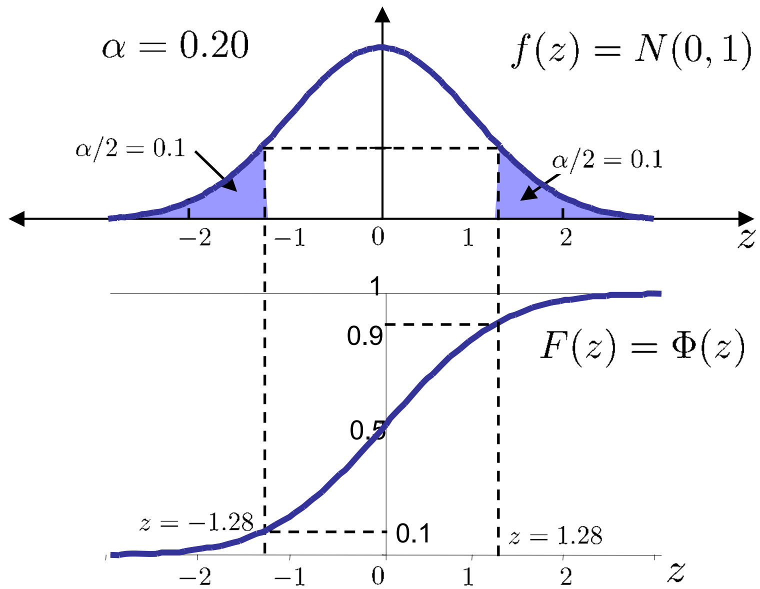
Confidence Intervals: Variance Known

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

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

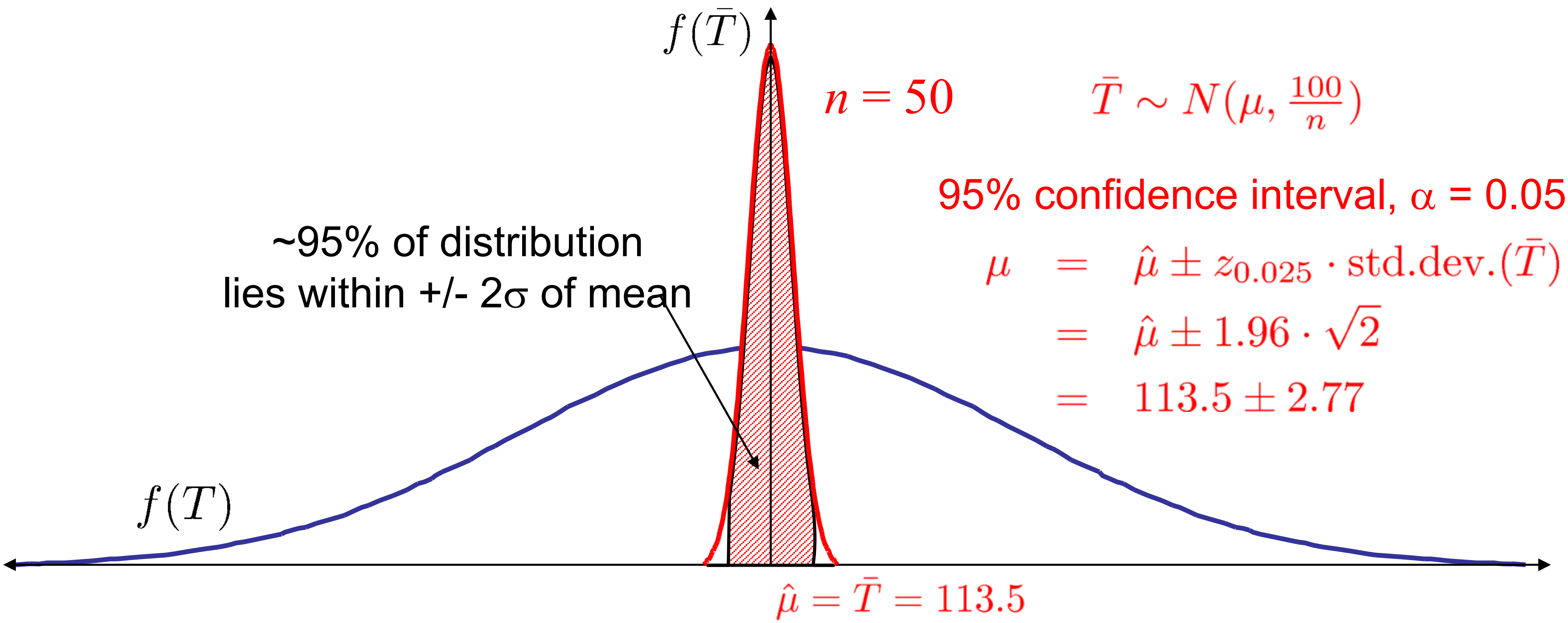
Remember the unit normal percentage points

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?



Reasoning & Sampling Distributions

- Example shows that we need to know our sampling distribution in order to reason about the sample and population parameters
- Other important sampling distributions:
  - Student's-t
    - Use instead of normal distribution when we don't know actual variation or σ
  - Chi-square
    - Use when we are asking about variances
  - F
    - Use when we are asking about ratios of variances

Sampling: The Chi-Square Distribution

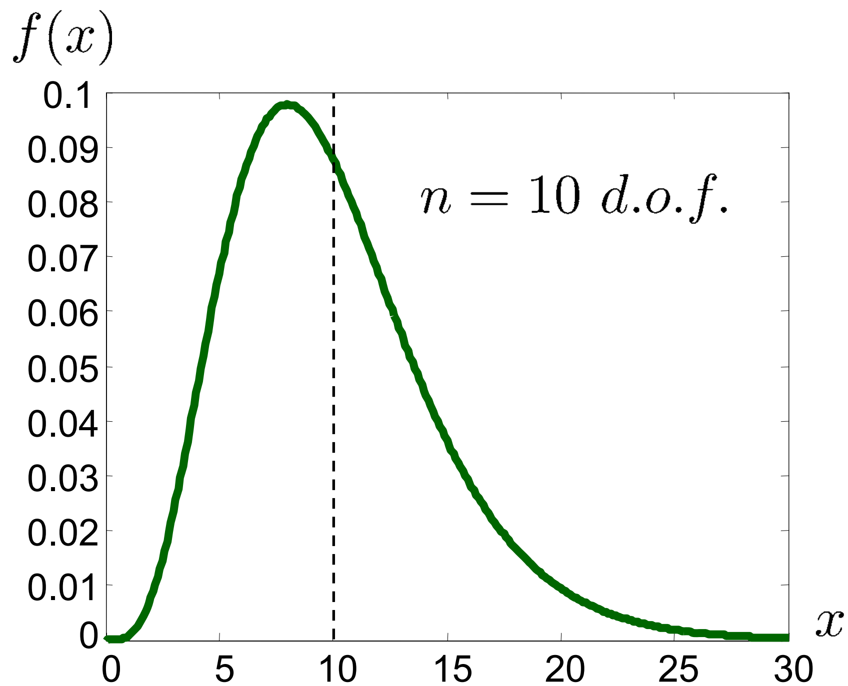
If $x_i \sim N(0,1)$ for $i = 1, 2, ..., n$ and $y = x_1^2 + x_2^2 + \cdots + x_n^2$, then $y \sim \chi_n^2$ or chi-square with n degrees of freedom.

- Typical use: find distribution of variance when mean is known
- Ex:

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

$$\frac{(n-1)s^2}{\sigma^2} \sim \chi_{n-1}^2$$

So if we calculate s^2, we can use knowledge of chi-square distribution to put bounds on where we believe the actual (population) variance sits

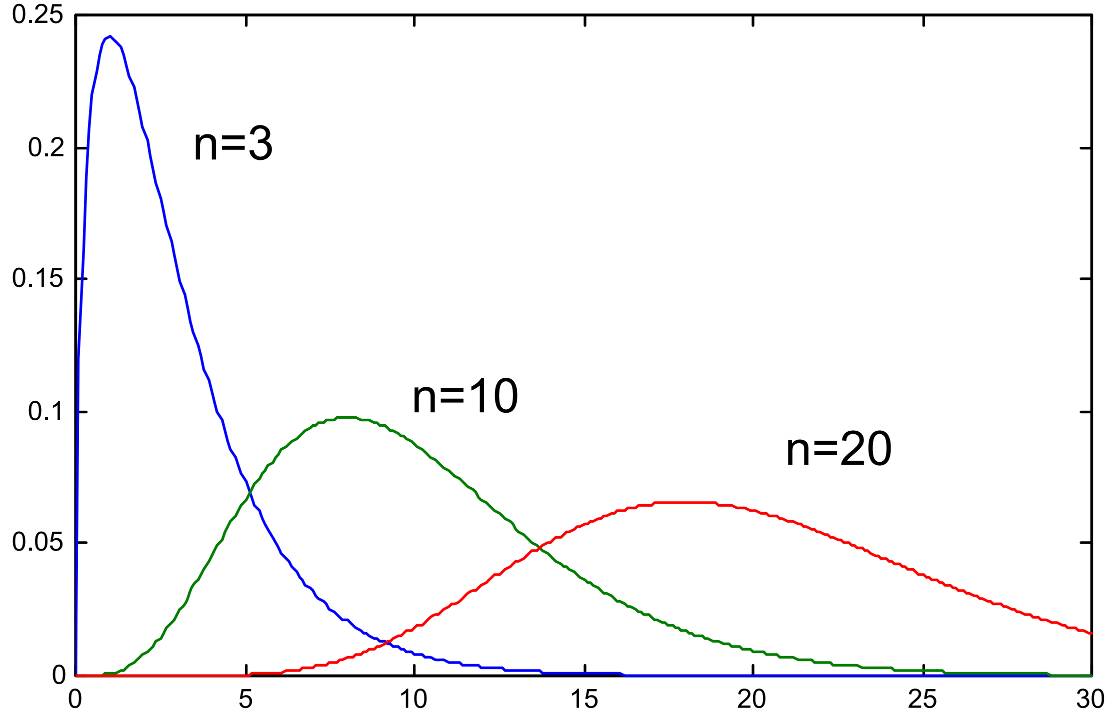


Note: $\mathrm{E}(\chi_n^2) = n$

Sampling: The Chi-Square Distribution

$$\frac{(n-1)s^2}{\sigma^2} \sim \chi^2_{n-1}$$

$$\mathrm{E}(\chi^2_n) = n$$



Sampling: The Student's-t Distribution

If $z \sim N(0,1)$ then $\frac{z}{y/k} \sim t_k$ with $y \sim \chi^2_k$ is distributed as a student t with k degrees of freedom.

- Typical use: Find distribution of average $\bar{x}$ <u>when σ is NOT known</u>

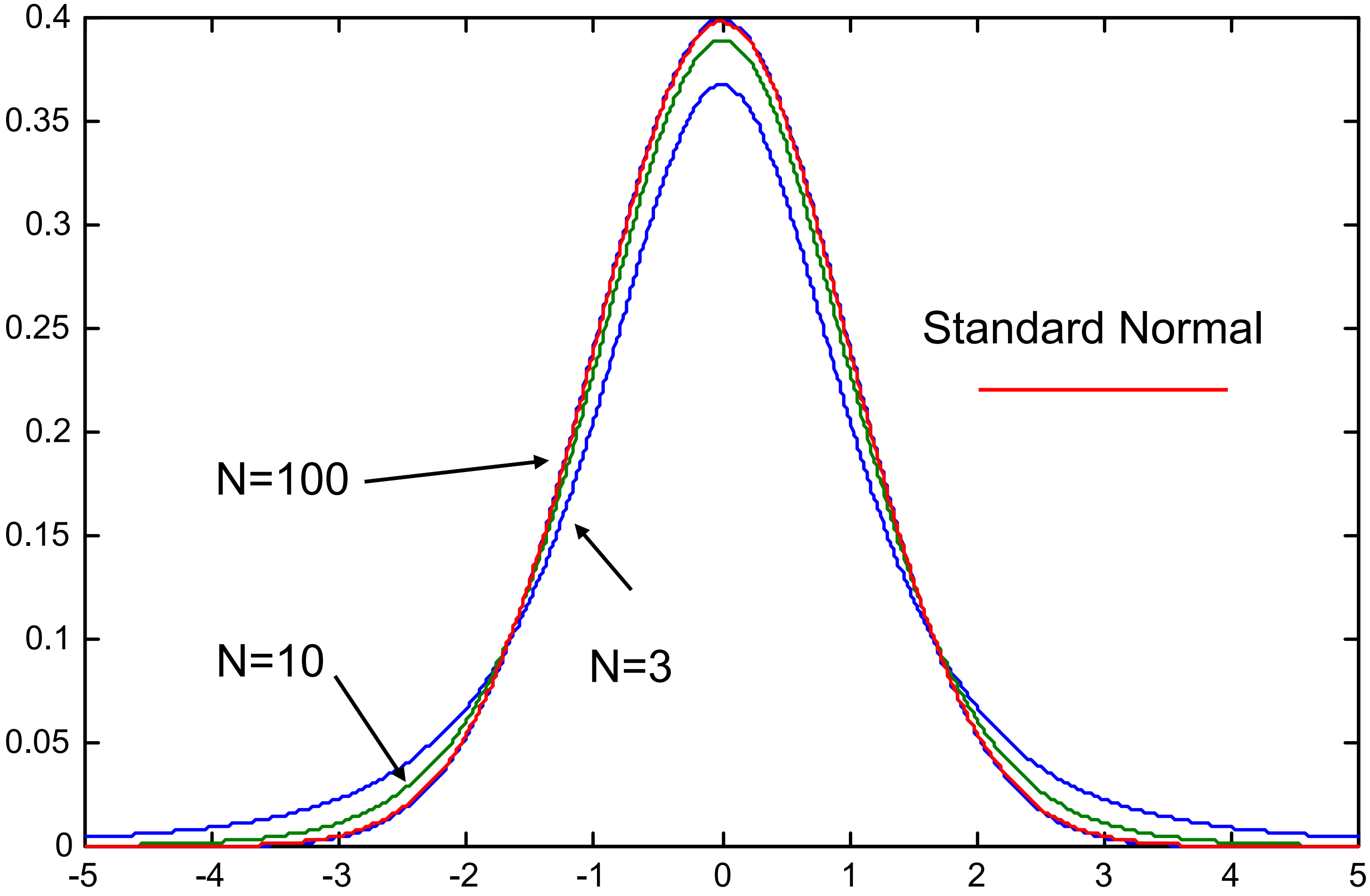
- Consider $x \sim N(\mu, \sigma^2)$. Then $\frac{\bar{x}-\mu}{\sigma/\sqrt{n}} \sim N(0,1)$

$$\frac{\bar{x}-\mu}{s/\sqrt{n}} = \frac{\frac{\bar{x}-\mu}{\sigma/\sqrt{2}}}{s/\sigma} \sim \frac{N(0,1)}{\sqrt{\frac{1}{n-1}\chi^2_{n-1}}} \sim t_{n-1}$$

- This is just the "normalized" distance from mean (normalized to our estimate of the sample variance)

Manufacturing

Sampling: The Student-t Distribution



Back to Our Example

- Suppose we do not know either the variance or the mean in our parts population:

$$T \sim N(\mu, \sigma^2) = N(\mu_{\text{unknown}}, \sigma^2_{\text{unknown}})$$

- We take our sample of size *n* = 50, and calculate

$$\bar{T} = \frac{1}{50}\sum_{i}^{50} T_i = 113.5 \qquad s_T^2 = \frac{1}{49}\sum_{i}^{50}(T_i - \bar{T})^2 = 102.3$$

- Best estimate of population mean and variance (std.dev.)?

$$\hat{\mu} = \bar{T} = 113.5 \qquad \hat{\sigma} = \sqrt{s_T^2} = 10.1$$

- If had to pick a range where μ would be 95% of time?

Have to use the appropriate sampling distribution:
In this case – the t-distribution (rather than normal distribution)

Confidence Intervals: Variance Unknown

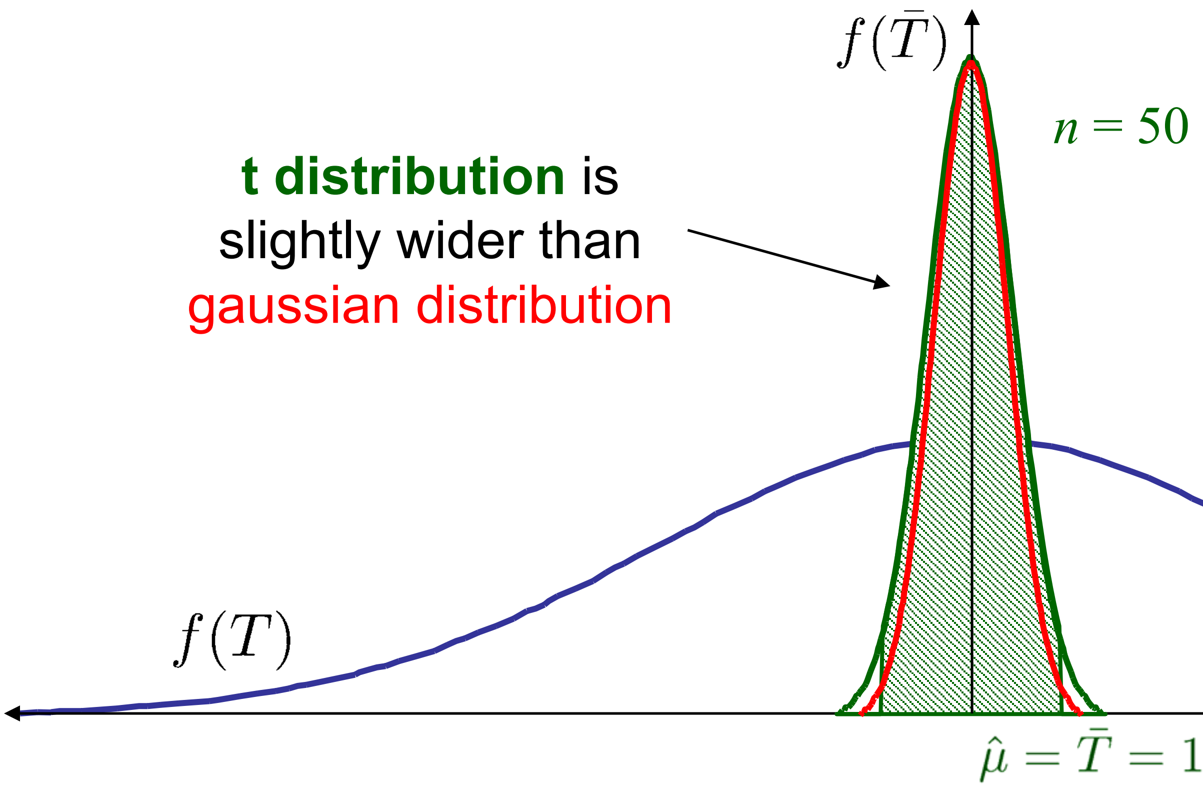
- Case where we don't know variance *a priori*
- Now we have to estimate not only the mean based on our data, but also estimate the variance
- Our estimate of the mean to some interval with (1-α)100% confidence becomes

$$\bar{x} - t_{\alpha/2,n-1} \cdot \frac{s}{\sqrt{n}} \;\leq\; \mu \;\leq\; \bar{x} + t_{\alpha/2,n-1} \cdot \frac{s}{\sqrt{n}}$$

Note that the t distribution is slightly wider than the normal distribution, so that our confidence interval on the true mean is not as tight as when we know the variance.

Example, Cont'd

- Third question: can we use knowledge of $\bar{T}$ distribution to reason about the actual (population) mean μ given observed (sample) mean – even though we weren't told σ?



$\bar{T} \sim t$ with $k = 49$ d.o.f.

95% confidence interval

$$\mu = \hat{\mu} \pm t_{0.025} \frac{s_T}{\sqrt{n}}$$
$$= 113.5 \pm 2.01 \cdot 10.1/\sqrt{50}$$
$$= 113.5 \pm 2.87$$

MIT Manufacturing

Once More to Our Example

- Fourth question: how about a confidence interval on our estimate of the **variance** of the thickness of our parts, based on our 50 observations?

Confidence Intervals: Estimate of Variance

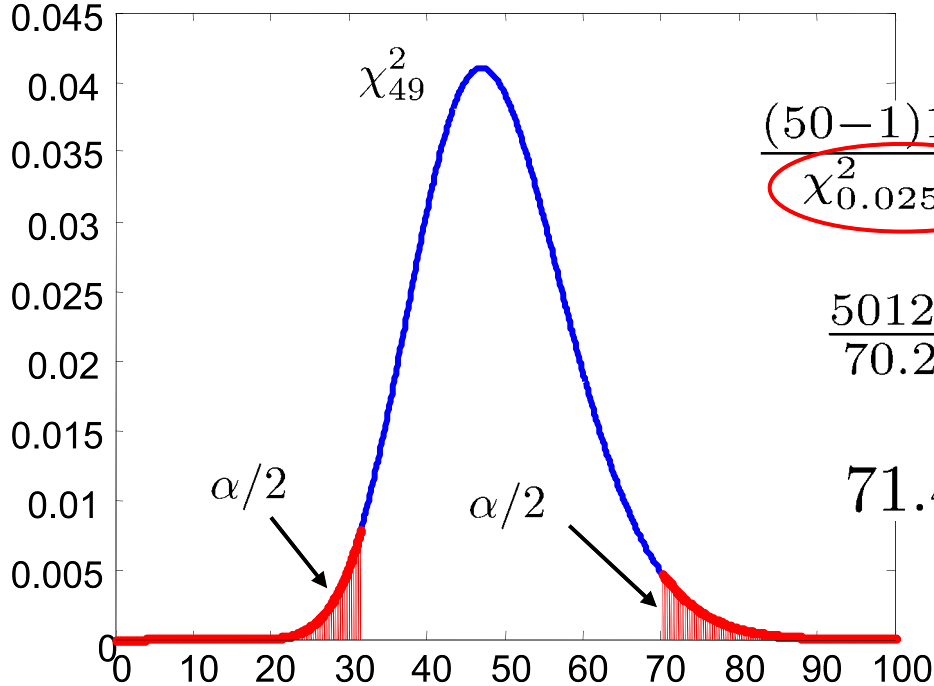
$$\frac{(n-1)s^2}{\chi^2_{\alpha/2,n-1}} \leq \sigma^2 \leq \frac{(n-1)s^2}{\chi^2_{1-\alpha/2,n-1}}$$

The appropriate sampling distribution is the Chi-square.
Because χ^2 is asymmetric, c.i. bounds not symmetric.

$$\frac{(n-1)s^2}{\sigma^2} \sim \chi^2_{n-1}$$

Example, Cont'd

- Fourth question: for our example (where we observed $s_T^2 = 102.3$) with $n = 50$ samples, what is the 95% confidence interval for the population variance?



$$\frac{(50-1)102.3}{\chi^2_{0.025,49}} \leq \sigma^2 \leq \frac{(50-1)102.3}{\chi^2_{0.975,49}}$$

$$\frac{5012.7}{70.22} \leq \sigma^2 \leq \frac{5012.7}{31.55}$$

$$71.4 \leq \sigma^2 \leq 158.1$$

Sampling: The F Distribution

If $y_1 \sim \chi^2_u$ and $y_2 \sim \chi^2_v$, then $R = \frac{y_1/u}{y_2/v} \sim F_{u,v}$ is an F distribution with u, v degrees of freedom.

- Typical use: compare the spread of two populations
- Example:
  - $x \sim N(\mu_x, \sigma^2_x)$ from which we sample $x_1, x_2, \ldots, x_n$
  - $y \sim N(\mu_y, \sigma^2_y)$ from which we sample $y_1, y_2, \ldots, y_m$
  - Then

$$\frac{s_x^2/\sigma_x^2}{s_y^2/\sigma_y^2} \sim F_{n-1,m-1} \quad \text{or} \quad \frac{\sigma_y^2}{\sigma_x^2} \sim \frac{s_x^2}{s_y^2} F_{n-1,m-1}$$

Concept of the F Distribution

- Assume we have a normally distributed population
- We generate two different random samples from the population
- In each case, we calculate a sample variance s_i^2
- What range will the ratio of these two variances take?
  F distribution
- Purely by chance (due to sampling) we get a range of ratios even though drawing from same population

Example:

- Assume $x \sim N(0,1)$
- Take 2 samples setss of size n = 20
- Calculate s_1^2 and s_2^2 and take ratio

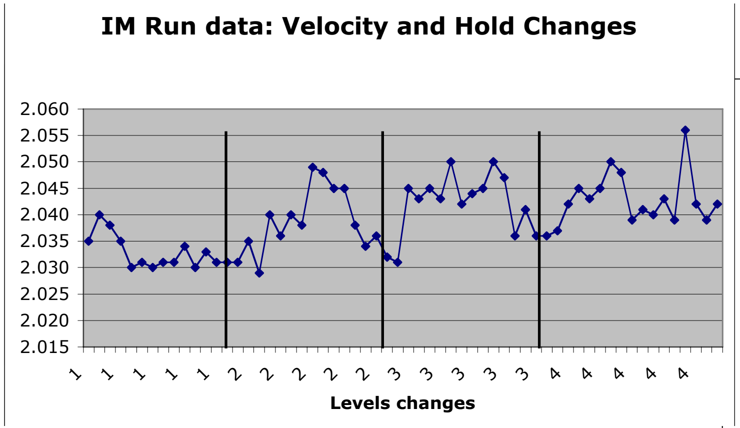
$$\frac{s_1^2}{s_2^2} \sim F_{19,19}$$

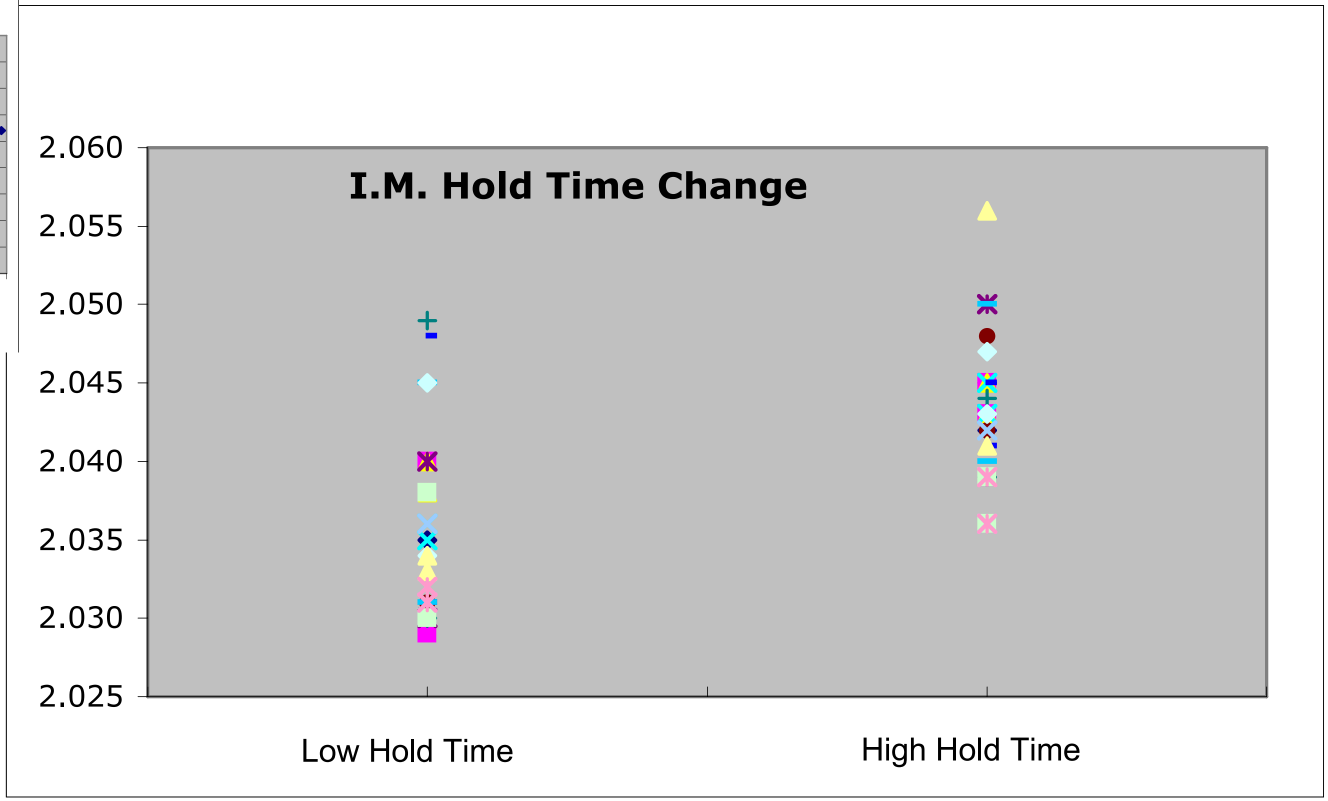
- 95% confidence interval on ratio

$$F_{\frac{\alpha}{2},19,19} = F_{0.025,19,19} = 2.53$$
$$F_{1-\frac{\alpha}{2},19,19} = F_{0.975,19,19} = 0.40$$

Large range in ratio!

Use of the F Distribution





Agenda

- Models for Random Processes
  - Probability Distributions & Random Variables
- Estimating Model Parameters with 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
- Next: Hypothesis tests

Statistical Inference and the Shewhart Hypothesis

- Statistical Hypotheses
  - Confidence of Predictions based on known or estimated pdf
- Relationship to Manufacturing Processes

Statistical Hypothesis

- e.g. hypothesize that mean has specific value
  – Based on Assumed Model (Distribution)
- Accept or reject hypothesis based on data and statistical model
  – i.e. based on degree of acceptable uncertainty or probability of error

Hypothesis Testing

- Given the hypothesis for the statistic ϕ *(e.g. the mean)*

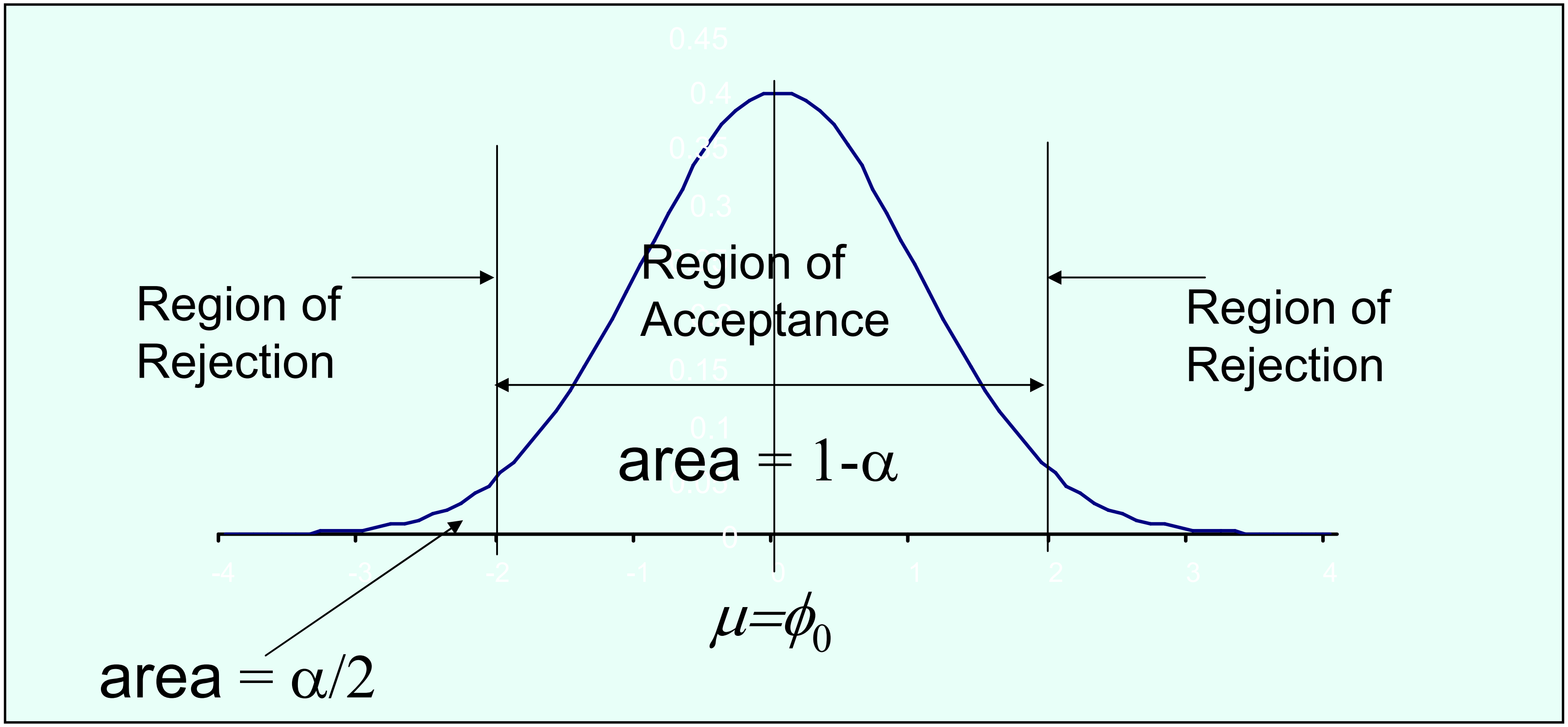
$$\mathrm{H}_0:\ \phi = \phi_0$$
$$\mathrm{H}_1:\ \phi \neq \phi_0$$

$$\left\{\begin{matrix}\mathrm{H}_0:\ \phi = \mu_0 \\ \mathrm{H}_1:\ \phi \neq \mu_0\end{matrix}\right\}$$

- No single sampled value $\hat{\phi}$ will equal ϕ_0
- How do we test the hypothesis given $\hat{\phi}$?
  - What is $p(\hat{\phi})$? (Sample Distribution?)
  - How well do we want to test H_o?
    - Significance
    - Confidence

The Test

- Assume a Distribution (e.g. $p(\hat{\phi})$ is Normal)



α is the *significance* of the test

Errors

- H_o is rejected when it is in fact true (Type I) (Significance)
  - p = ?

$p=\alpha$ for two sided and $\alpha/2$ for one sided tests

- H_o is accepted when it is in fact false (Type II)
  - p = ?

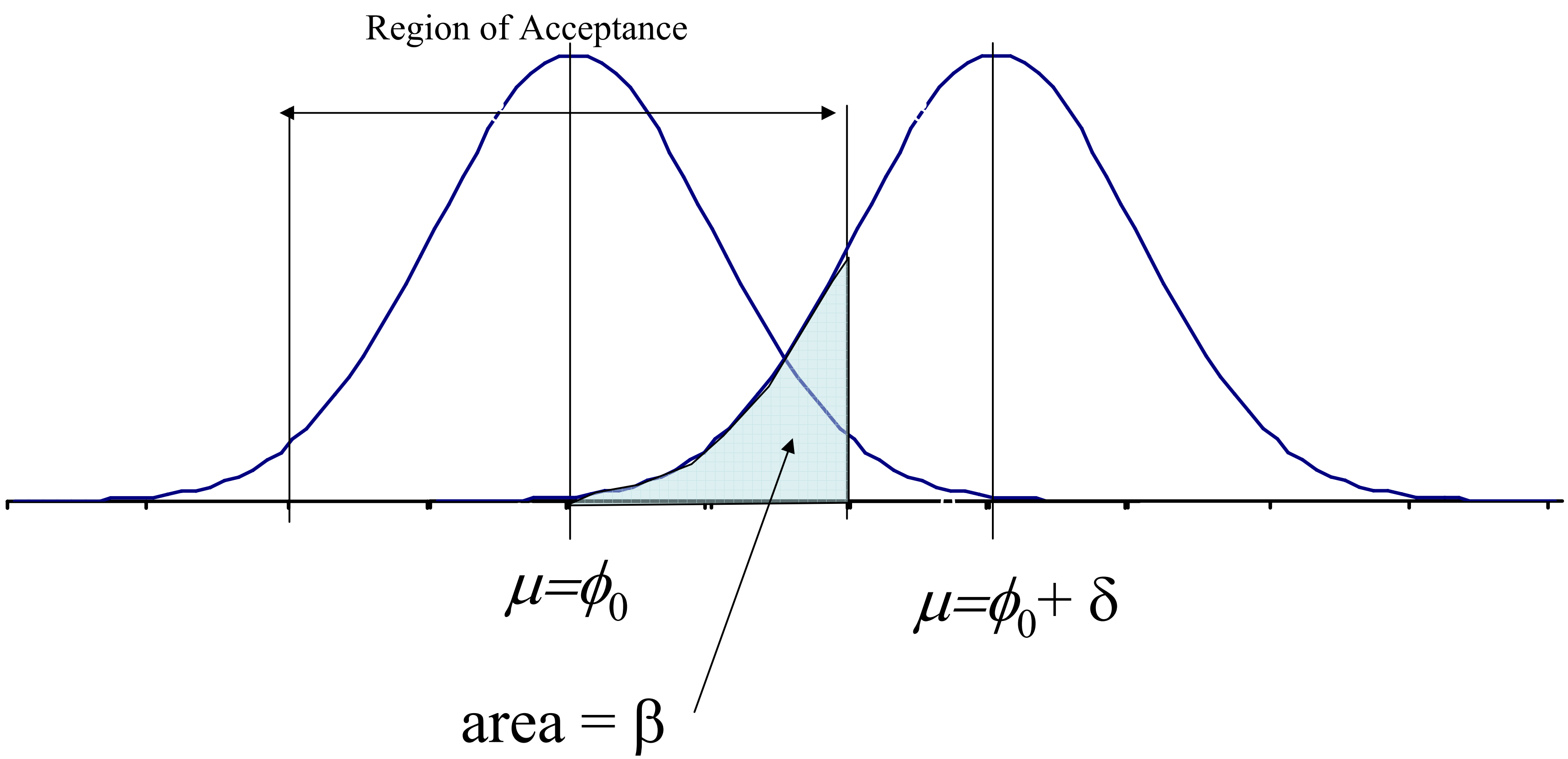
$=\beta$

… What is β ?

Type II Errors

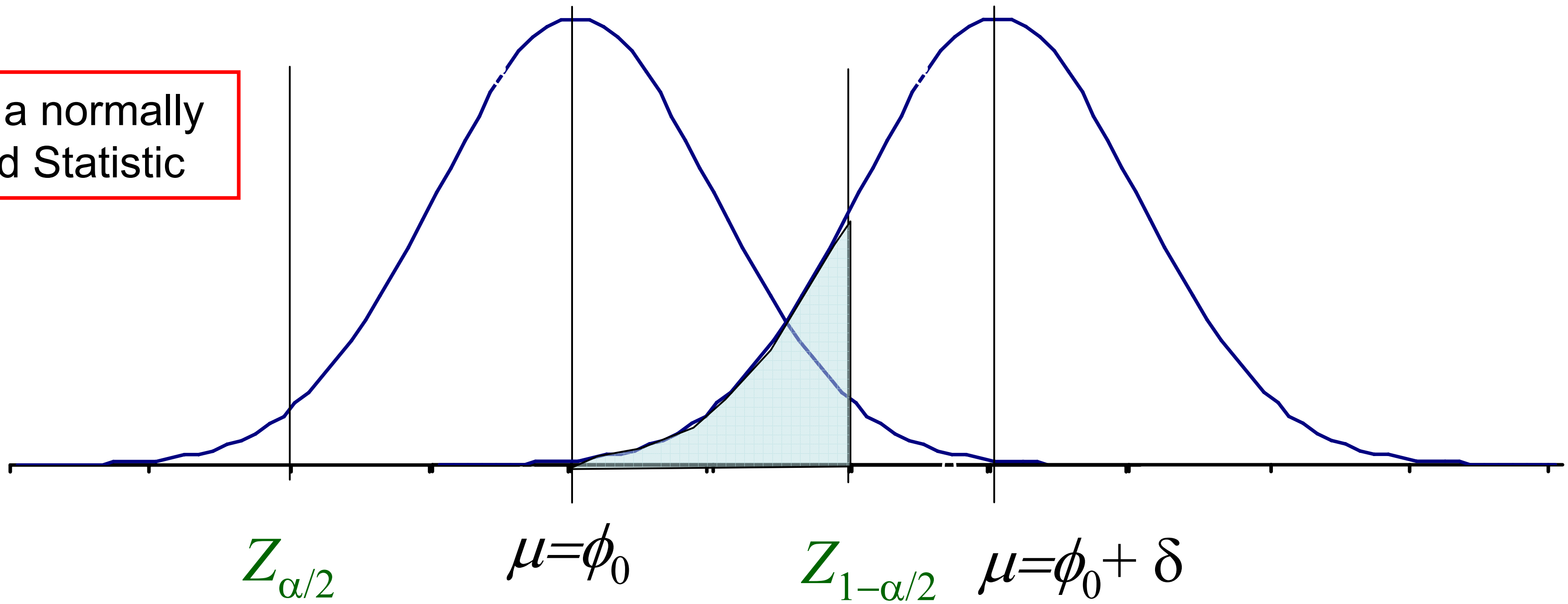
- Assume a shift in the true distribution $p(\hat{\phi})$ of δ
- Assess the probability that we fall in the acceptance region after a shift of δ occurred

Type II Errors



Calculating β

Assuming a normally distributed Statistic



$$\beta = \Phi(Z_{1-\alpha/2} - \Delta) - \Phi(Z_{\alpha/2} - \Delta)$$

$$\Delta = \frac{\delta}{\sigma/\sqrt{n}}$$

Normalized deviation

Manufacturing

Applications

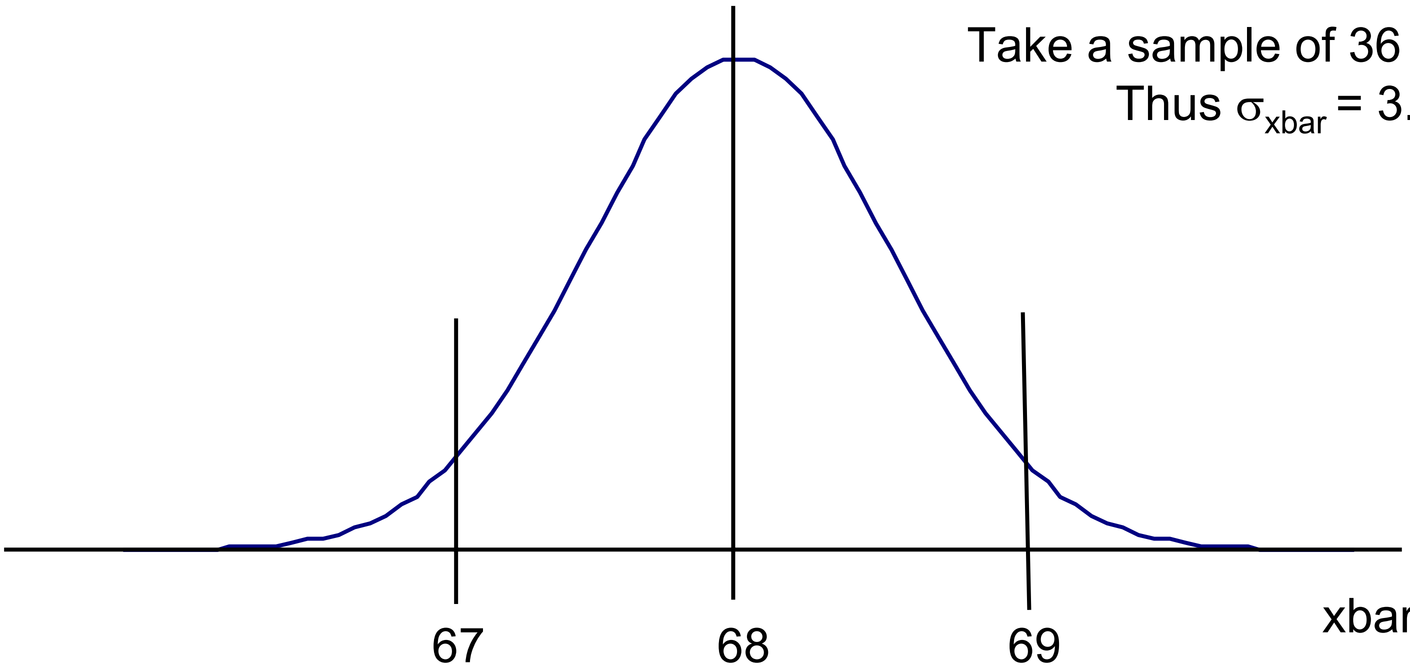
- Tests on the Mean
  - Is the mean of “new” data the same as prior data (I.e from the same distribution?)

  Or
  - Did a significant change occur?
- Variances of a Population
  - Is the variance of “new” data the same as prior data (i.e from the same distribution?)
- What are the “parent distributions” if we only have sample data?
  - Sample distributions

Example: Average Weight

- Hypothesize that average weight of a population is 68 kg and σ=3.6
  - H_0: $\mu = 68$
  - H_1: $\mu \neq 68$
  - Assume an acceptance region of +- 1 kg
  - What is α or significance of test?
    - Probability of a type I error
  - What is β
    - Probability of type II error

Significance from Interval



Take a sample of 36 weights for xbar
Thus $\sigma_{xbar} = 3.6/6 = 0.6$

$$z_1 = \frac{(67-68)}{0.6} = -1.67 \qquad z_2 = \frac{(69-68)}{0.6} = +1.67$$

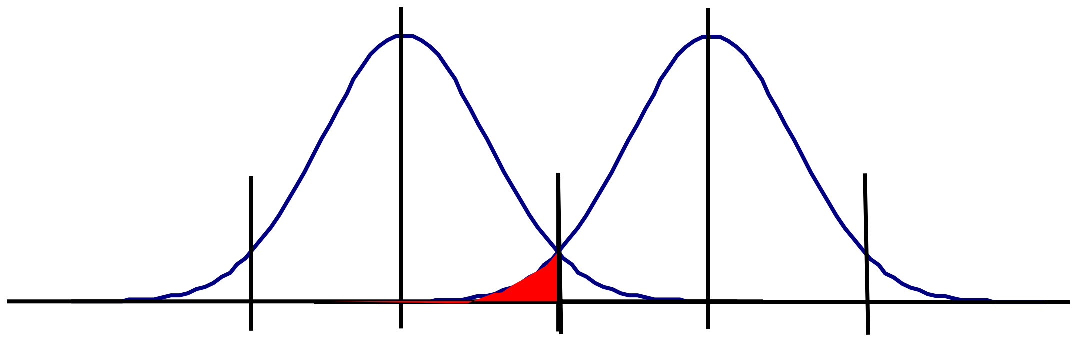
$$\alpha = P(Z < -1.67) + P(Z > 1.76) = 2P(Z < -1.67) = 0.095$$

9.5% chance of rejecting H_0 even if true

Effect of increasing range?
Effect of increasing n?

β Error

Assume we must reject H_0 if $\mu<66$ or $\mu>70$



i.e. $\beta = P(67 \le xbar \le 69)$ when $\mu = 70$

$$\beta = P(-6.67 \le Z \le -2.22) = 0.0132$$

1.3% chance of accepting H_o when it is false

From symmetry - same result for μ = 66

Effect of increasing range?
Effect of increasing n?

Operating Characteristic Curve Dependence of α and β

Note that the expression for β: depends on α, *n* and δ

From Montgomery "Introduction to Statistical Quality Control, 4th ed. 2000

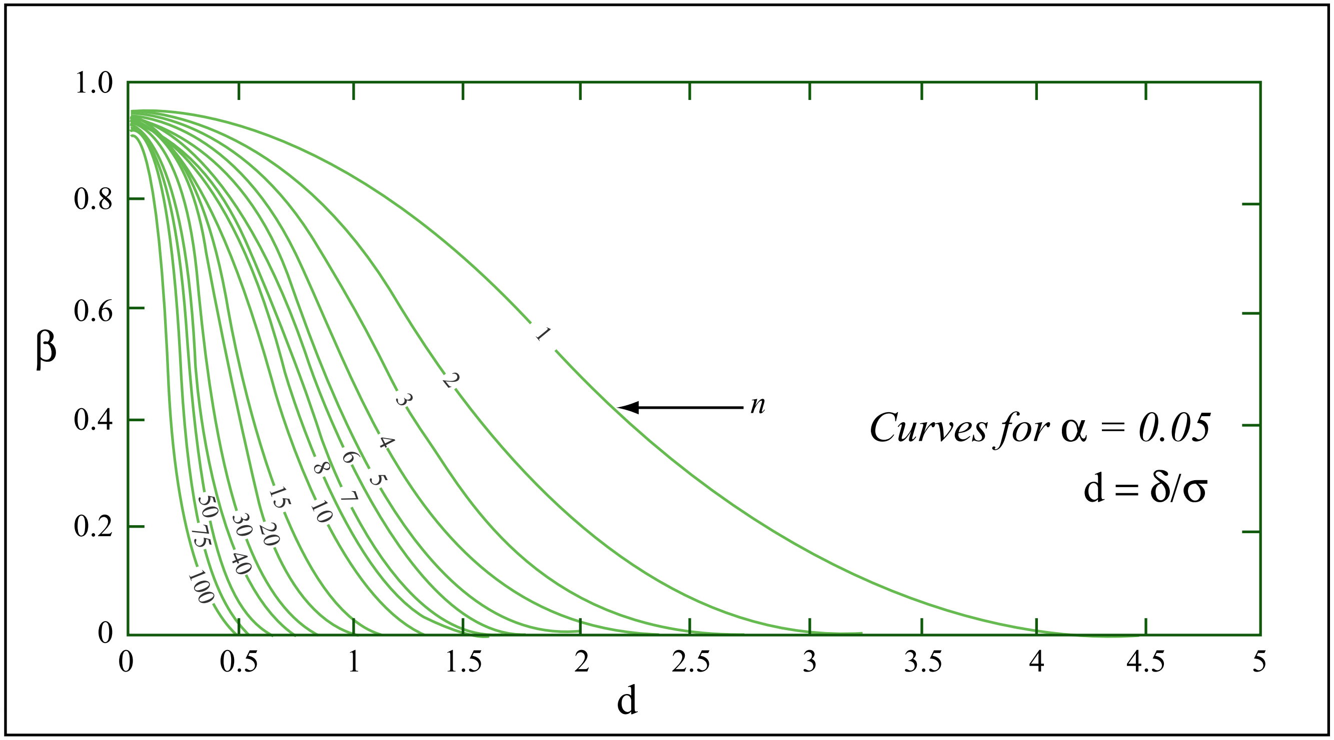


Figure by MIT OpenCourseWare.

Manufacturing

Some Typical Hypothesis

- Inference about Variance from Samples
  - Test Statistic?
  - Which Distribution to Use ?
- Inference about Mean
  - Knowing σ
  - Not knowing σ

Summary

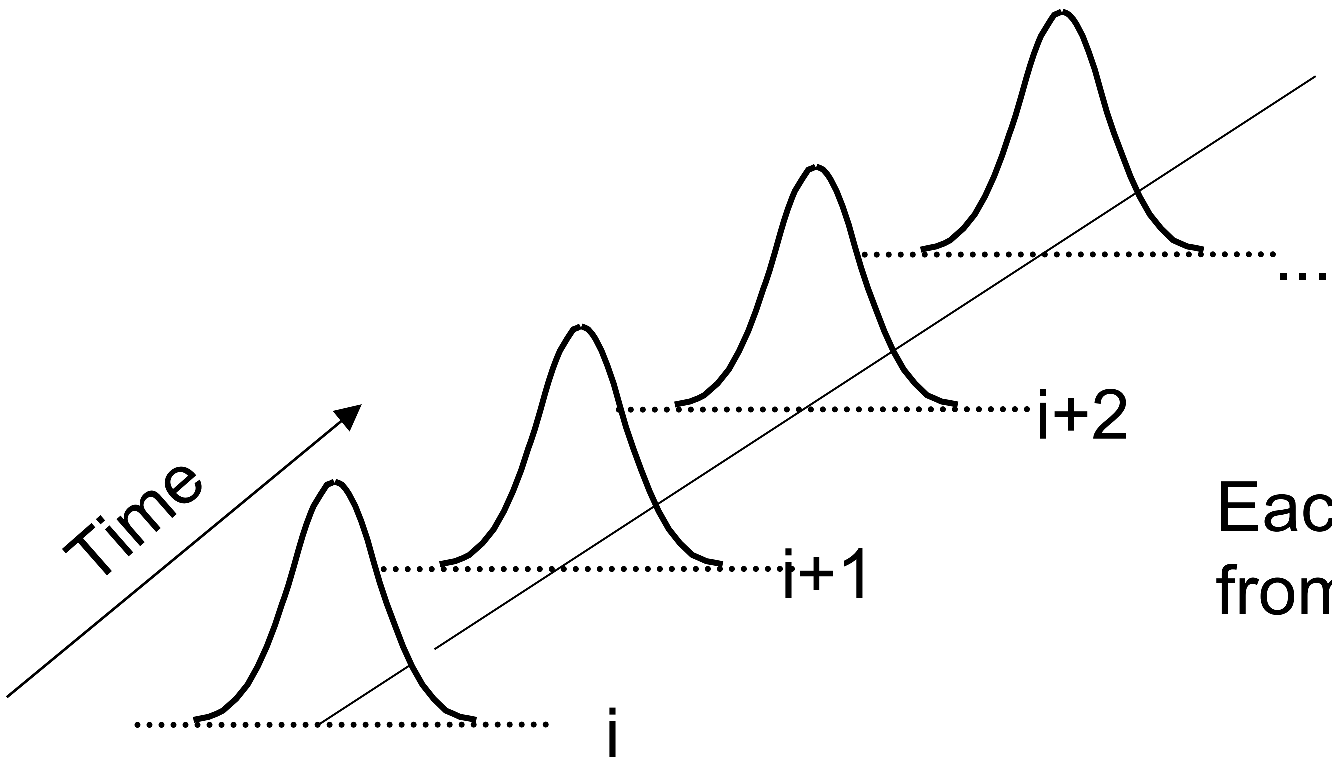
- Pick Significance Level α
- Determine an acceptable β
  - (1-β) is call the “power” of the test

- What is the effect of the number of samples (n)?

On to *Process Control*

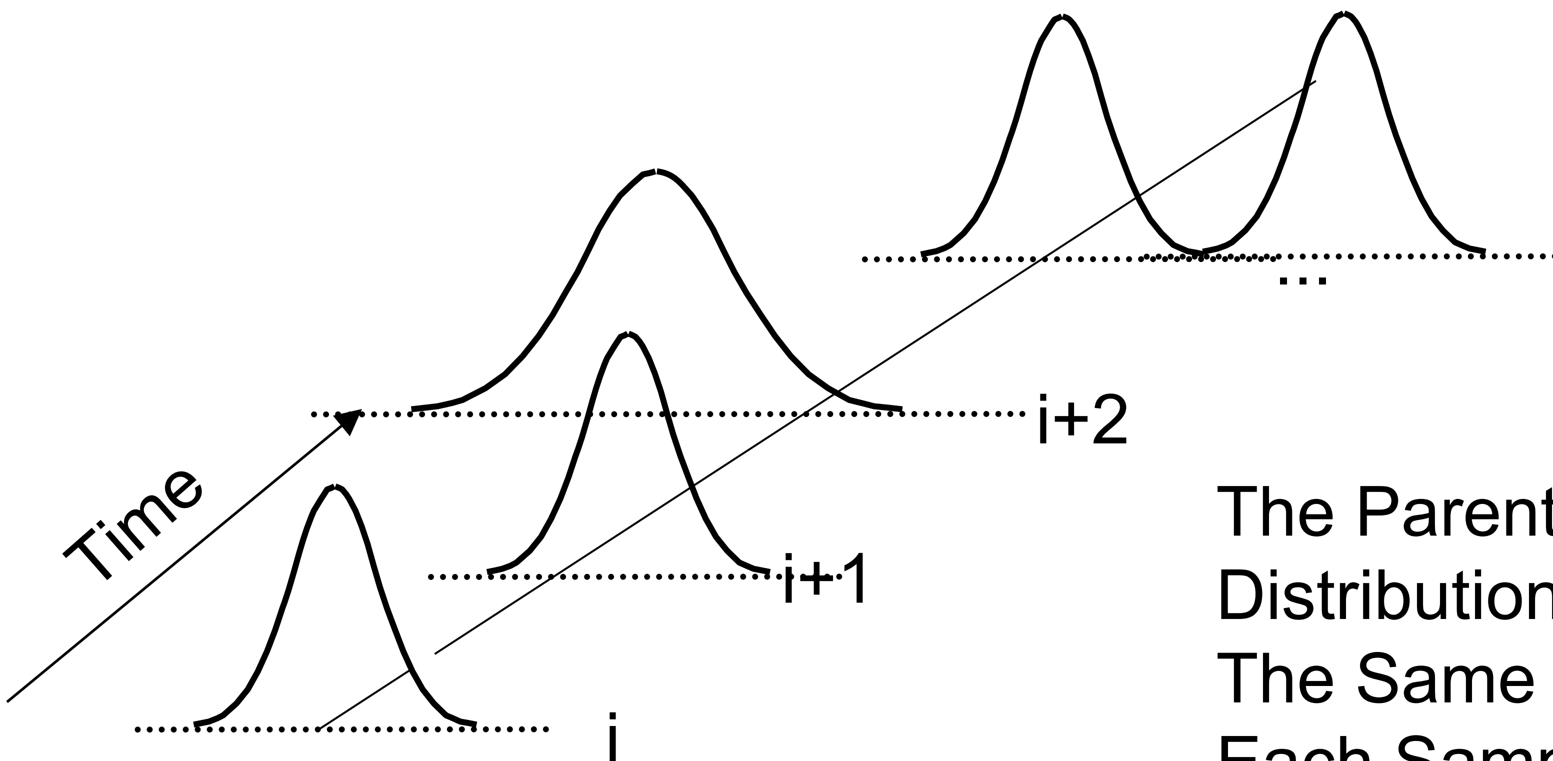
- How does all this relate to our problem?
- What assumptions must we make?
- What statistical tests should we use?
- What are the best procedures to use in a production environment?

"In-Control"



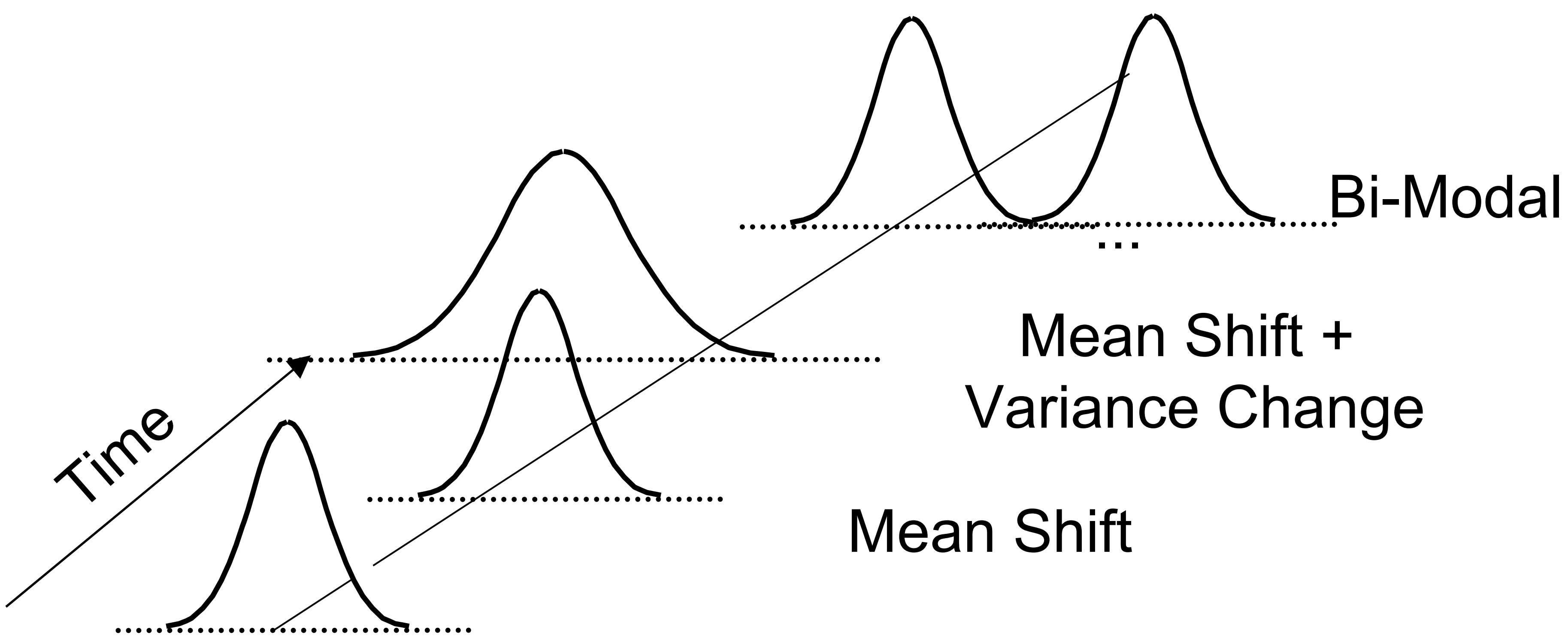
Each Sample is from Same Parent

"Not In-Control"



The Parent Distribution is Not The Same at Each Sample

"Not In-Control"



Xbar and S Charts

- Shewhart:
  - Plot *sequential average* of process
    - Xbar chart
    - Distribution?
  - Plot sequential sample standard deviation
    - S chart
    - Distribution?

Conclusions

- Hypothesis Testing
  - Use knowledge of PDFs to evaluate hypotheses
  - Quantify the degree of certainty (α and β)
  - Evaluate effect of sampling and sample size
- Shewhart Charts
  - Application of Statistics to Production
  - Plot Evolution of Sample Statistics $\bar{x}$ and S
  - Look for Deviations from Model