

SMA 6304 / MIT 2.853 / MIT 2.854
Manufacturing Systems

Lecture 9: Statistical Inference

Lecturer: Prof. Duane S. Boning

Agenda

1. Review: Probability Distributions & Random Variables
2. **Sampling:** Key distributions arising in sampling
 - Chi-square, t, and F distributions
3. **Estimation:**
Reasoning about the population based on a sample
4. Some basic confidence intervals
 - Estimate of mean with variance known
 - Estimate of mean with variance not known
 - Estimate of variance
5. Hypothesis tests

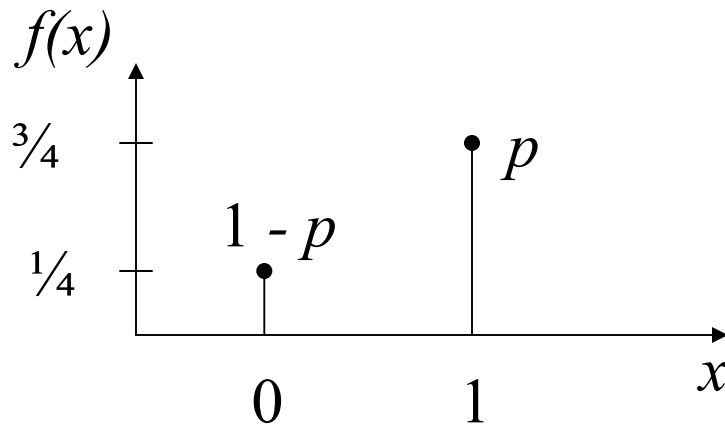
Discrete Distribution: Bernoulli

- Bernoulli trial: an experiment with two outcomes

$$\Pr(\text{success}) = \Pr(1)$$

$$\Pr(\text{failure}) = \Pr(0)$$

- Probability mass function (pmf): $f(x, p) = \begin{cases} p & x = 1 \\ 1 - p & x = 0 \end{cases}$



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

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

Discrete Distribution: Binomial

- Repeated random Bernoulli trials

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

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 binomial

- 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

Discrete Distribution: Poisson

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

- Mean: $\mu = \lambda$
- Variance: $\sigma^2 = \lambda$
- Example applications:
 - # misprints on page(s) of a book
 - # transistors which fail on first day of operation
- Poisson is a good approximation to Binomial when n is large and p is small (< 0.1)

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

Continuous Distributions

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

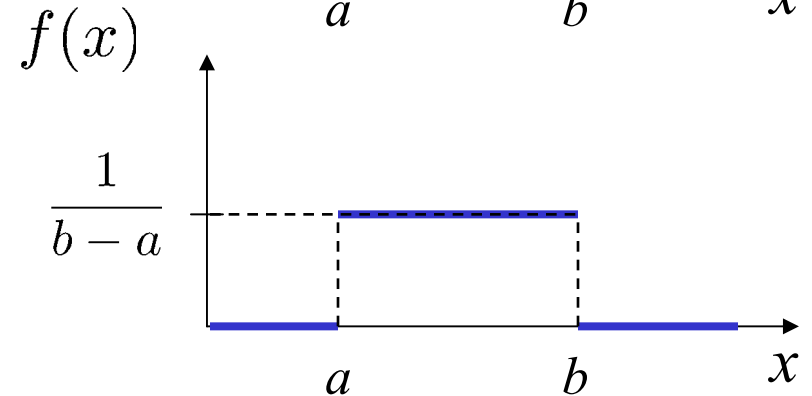
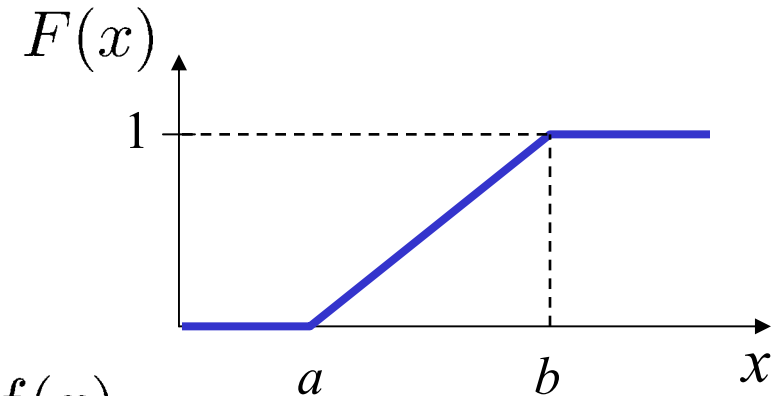
Continuous Distribution: Uniform

- cdf

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

- pdf

$$f(x) = \begin{cases} \frac{1}{b-a} & a \leq x < b \\ 0 & \text{otherwise} \end{cases}$$



$$x \sim U(a, b)$$

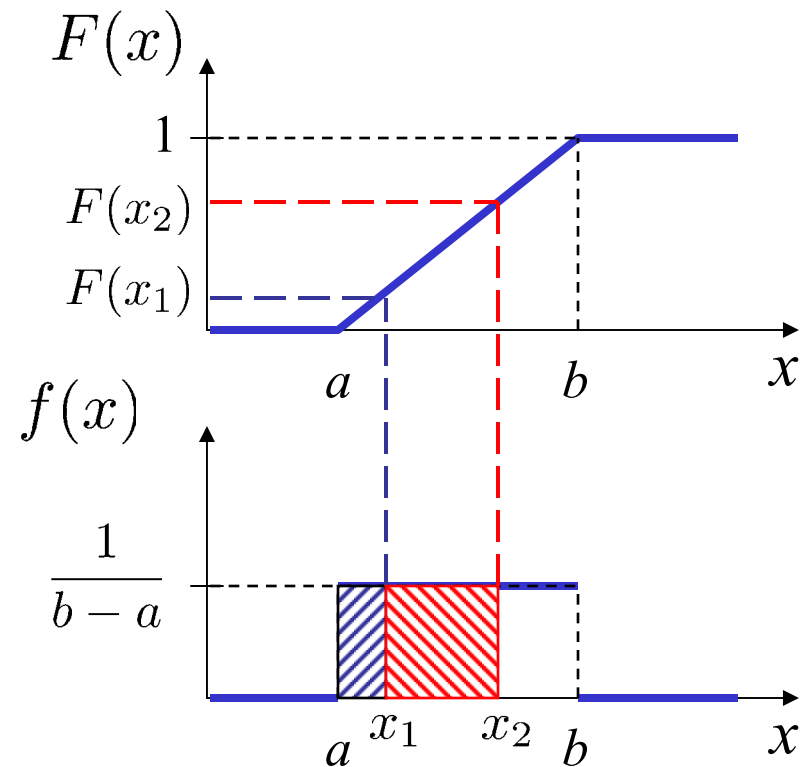
Standard Questions You Should Be Able To Answer (For a Known cdf or pdf)

- Probability x less than or equal to some value

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

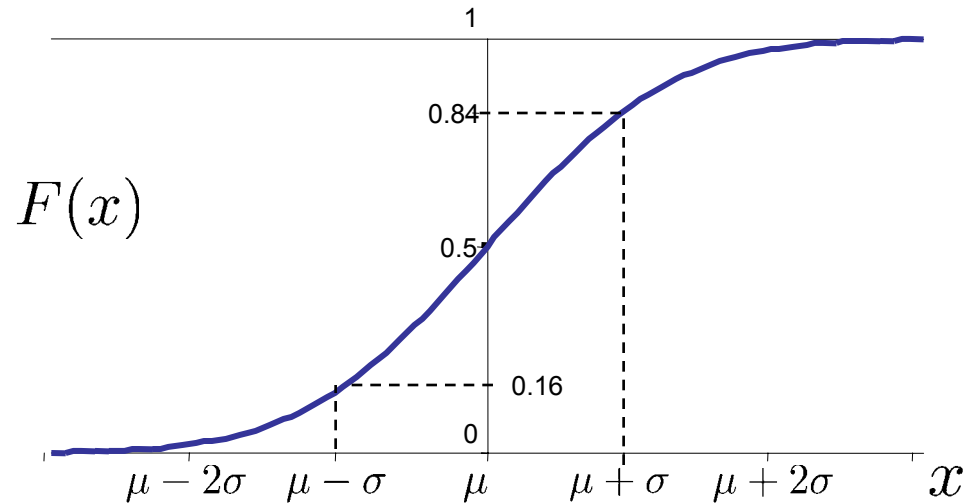
- Probability x sits within some range

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



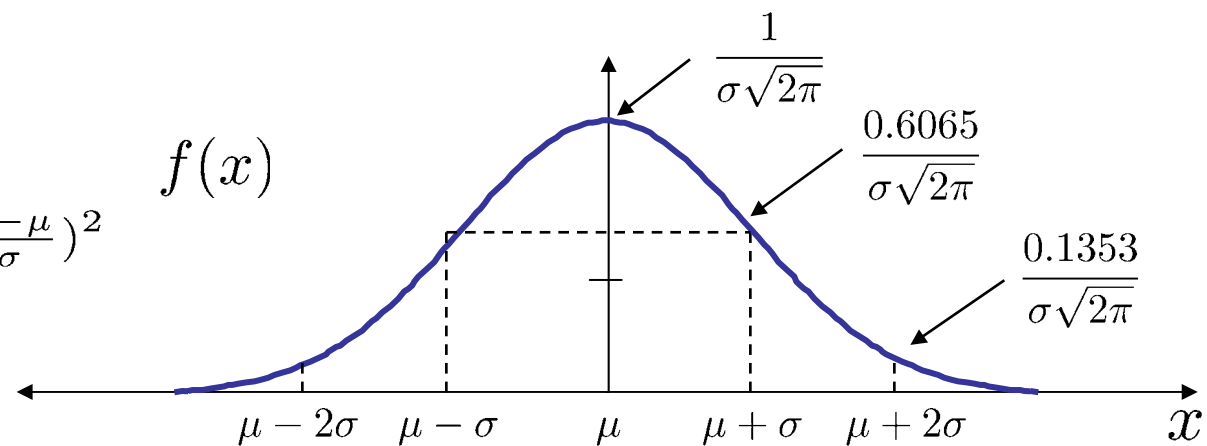
Continuous Distribution: Normal (Gaussian)

- cdf



- pdf $x \sim N(\mu, \sigma^2)$

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



Continuous Distribution: Unit Normal

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

- Mean $E(z) = 0$

- Variance $\text{Var}(z) = 1 \Rightarrow \text{std.dev.}(z) = 1$

- 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

$$\Phi(z) = \alpha/2$$

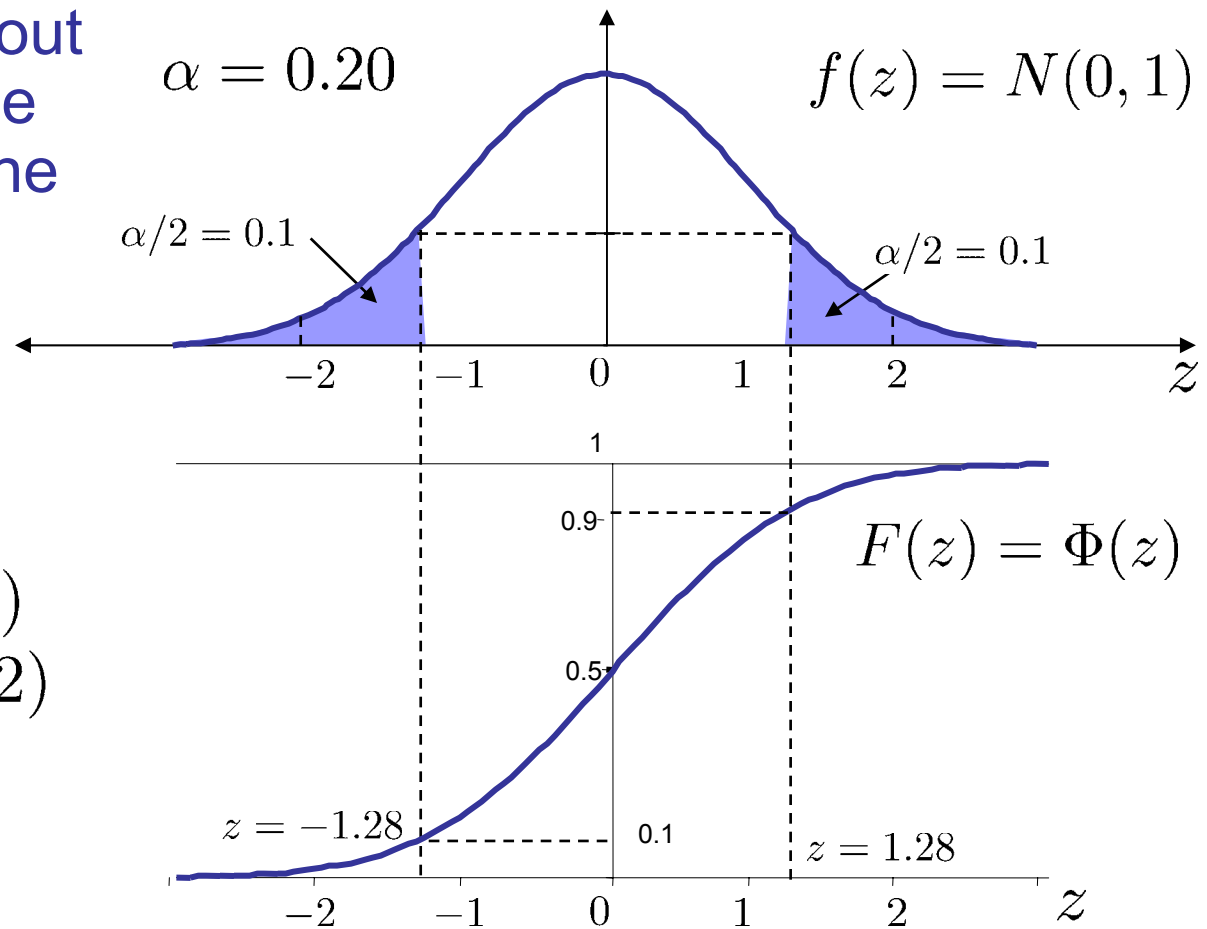
$$1 - \Phi(z) = 1 - \alpha/2$$

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

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

$$z_{0.10} = -1.28$$

$$z_{0.90} = 1.28$$



Philosophy

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

- Beliefs:
 - Distribution or model form
 - Distribution/model parameters
- Evidence:
 - Finite set of observations or data drawn from a population
- Models:
 - Seek to explain data

Moments of the Population vs. Sample Statistics

Population

Sample

- Mean

$$\mu = \mu_x = E(x)$$

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

- Variance

$$\sigma^2 = \sigma_{xx}^2 = E[(x - \mu_x)^2]$$

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

- Standard Deviation

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

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

- Covariance

$$\begin{aligned} \sigma_{xy}^2 &= E[(x - \mu_x)(y - \mu_y)] \\ &= E(xy) - E(x)E(y) \end{aligned}$$

$$s_{xy}^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})$$

- Correlation Coefficient

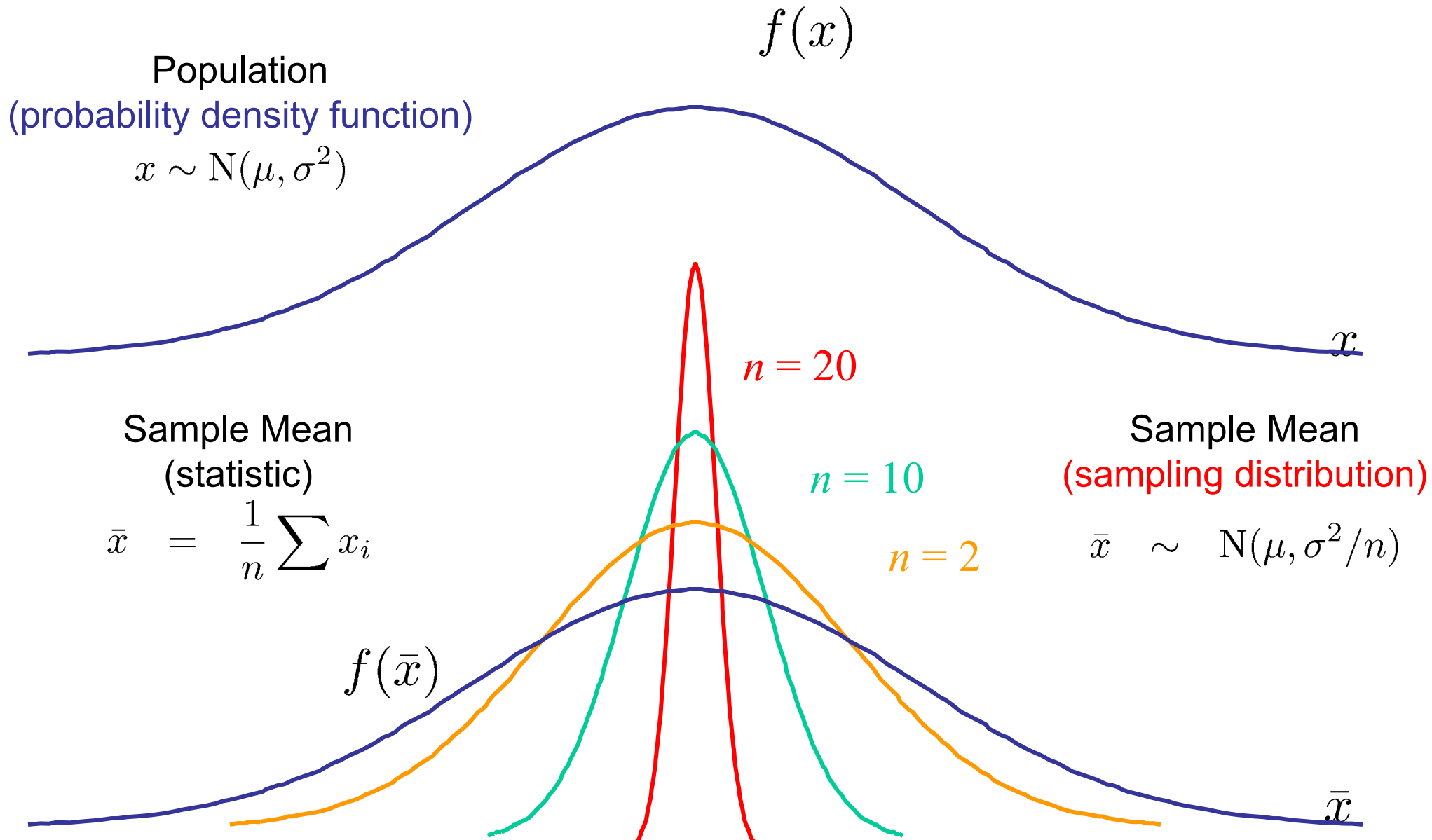
$$\rho_{xy} = \frac{\sigma_{xy}}{\sigma_x \sigma_y} = \frac{\text{Cov}(xy)}{\sqrt{\text{Var}(x)\text{Var}(y)}}$$

$$r_{xy} = \frac{s_{xy}}{s_x s_y}$$

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

Population vs. Sampling Distribution



Sampling and Estimation, cont.

- Sampling
- Random sampling
- Statistic
- A **statistic** is a random variable, which itself has a **sampling 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 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?

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 \bar{T} ?

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

$$\begin{aligned} E(\bar{T}) &= \mu \\ \text{Var}(\bar{T}) &= \sigma^2/n = 100/50 \\ &\text{Normally distributed} \end{aligned}$$

- 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

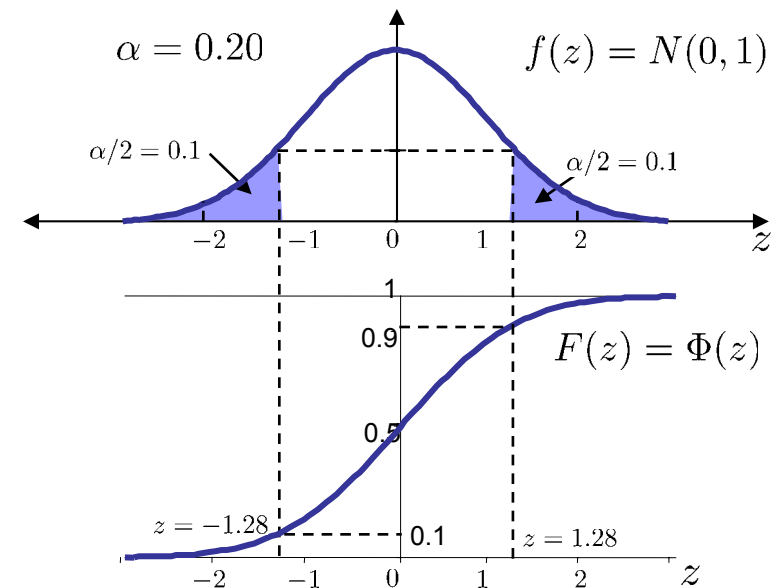
- Point Estimation:
 - Find best values for parameters of a distribution
 - Should be
 - Unbiased: expected value of estimate should be true value
 - Minimum variance: should be estimator with smallest variance
- Interval Estimation:
 - Give bounds that contain actual value with a given probability
 - Must know sampling distribution!

Confidence Intervals: Variance Known

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

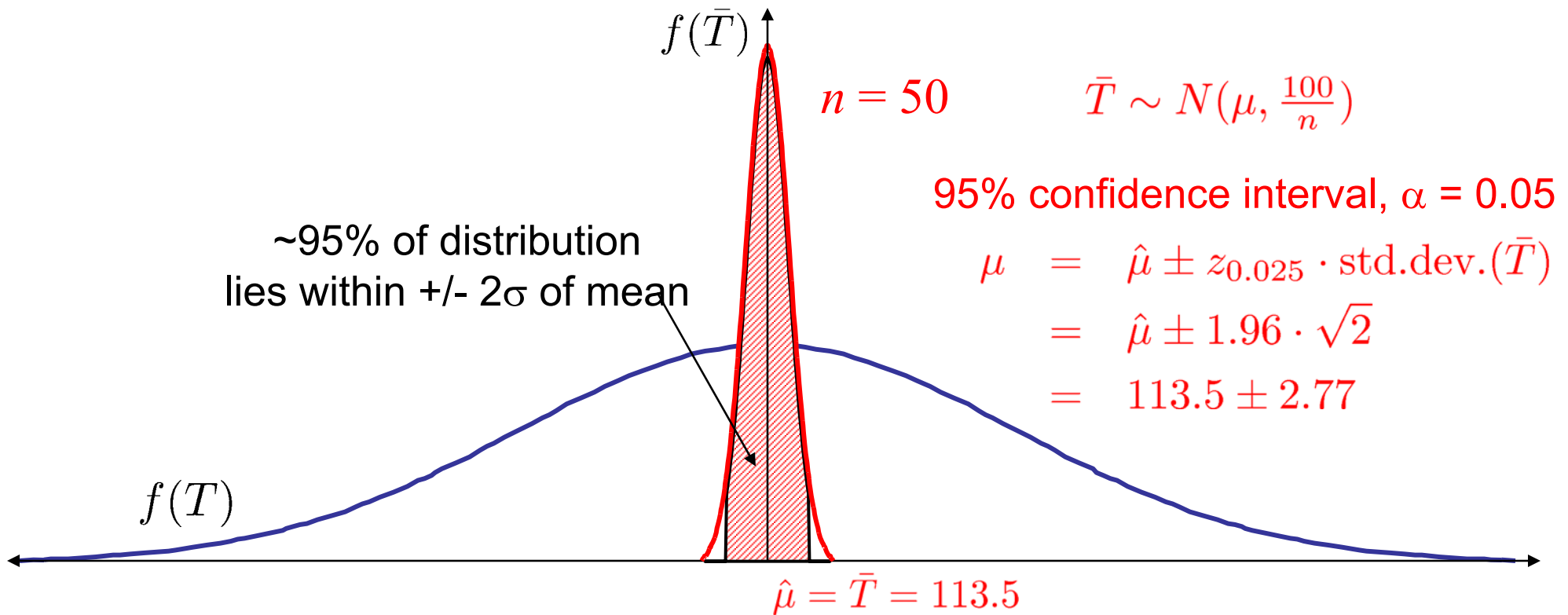
$$\bar{x} - z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{x} + z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}$$

- 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-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

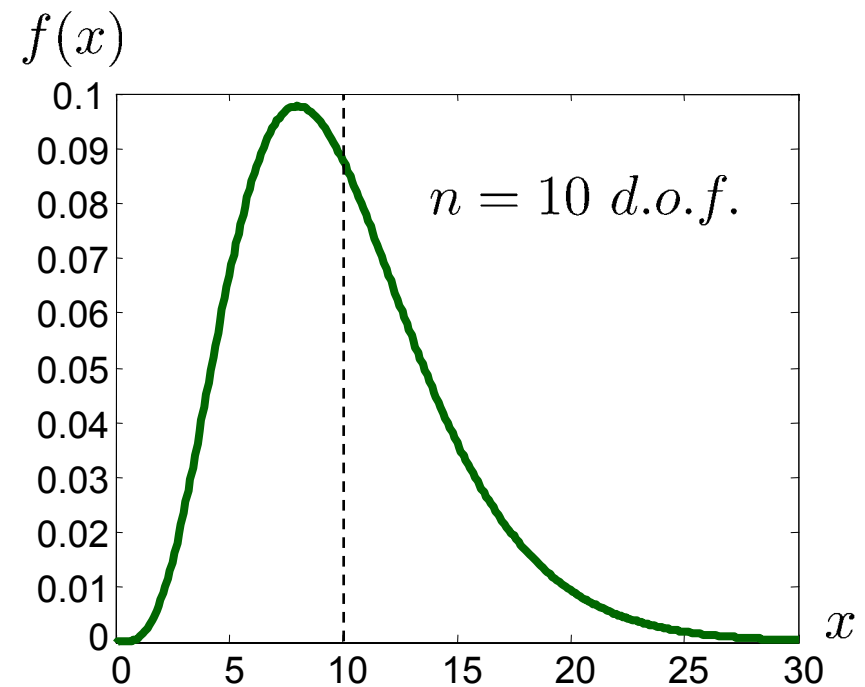
Sampling: The Chi-Square Distribution

If $x_i \sim N(0, 1)$ for $i = 1, 2, \dots, n$ and $y = x_1^2 + x_2^2 + \dots + x_n^2$, 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: $x_i \sim N(\mu, \sigma^2)$
$$\frac{(n-1)s^2}{\sigma^2} \sim \chi_{n-1}^2$$

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: $E(\chi_n^2) = n$

Sampling: The Student-t Distribution

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

- Typical use: Find distribution of average when σ is NOT known
- For $k \rightarrow \infty$, $t_k \rightarrow N(0, 1)$
- Consider $x_i \sim N(\mu, \sigma^2)$. Then

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

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

Back to our Example

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

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

- 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

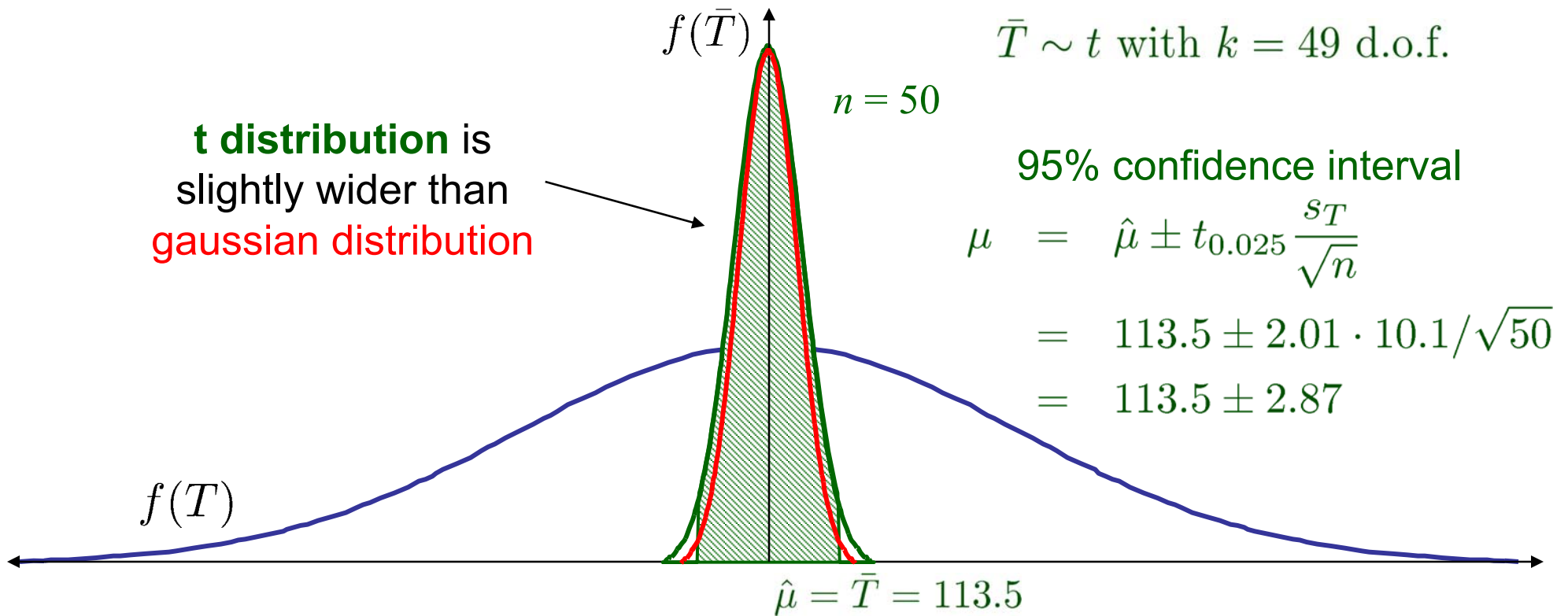
- Case where we don't know variance a priori
- Now we have to estimate not only the mean based on our data, but also estimate the variance
- Our estimate of the mean to some interval with $(1-\alpha)100\%$ confidence becomes

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

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 ?



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?

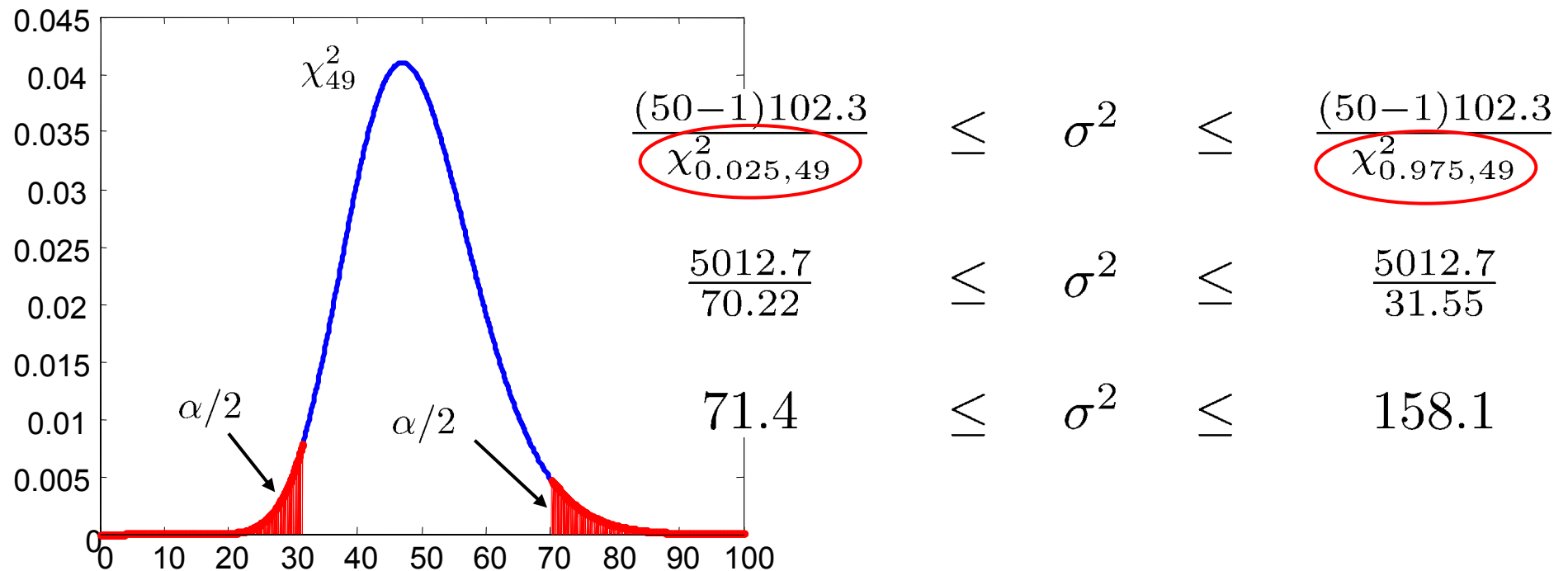
Confidence Intervals: Estimate of Variance

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

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

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?



Sampling: The F Distribution

If $y_1 \sim \chi_u^2$ and $y_2 \sim \chi_v^2$, 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:
 - $x \sim N(\mu_x, \sigma_x^2)$ from which we sample x_1, x_2, \dots, x_n
 - $y \sim N(\mu_y, \sigma_y^2)$ from which we sample y_1, y_2, \dots, 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 samples 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

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

Large range in ratio!

Hypothesis Testing

- A statistical hypothesis is a statement about the parameters of a probability distribution
- H_0 is the “null hypothesis”
 - E.g. $H_0 : \mu = \mu_0$
 - Would indicate that the machine is working correctly
- H_1 is the “alternative hypothesis”
 - E.g. $H_1 : \mu \neq \mu_0$
 - Indicates an undesirable change (mean shift) in the machine operation (perhaps a worn tool)
- In general, we formulate our hypothesis, generate a random sample, compute a statistic, and then seek to reject H_0 or fail to reject (accept) H_0 based on probabilities associated with the statistic and level of confidence we select

Which Population is Sample x From?

- Two error probabilities in decision:

- Type I error: “false alarm”

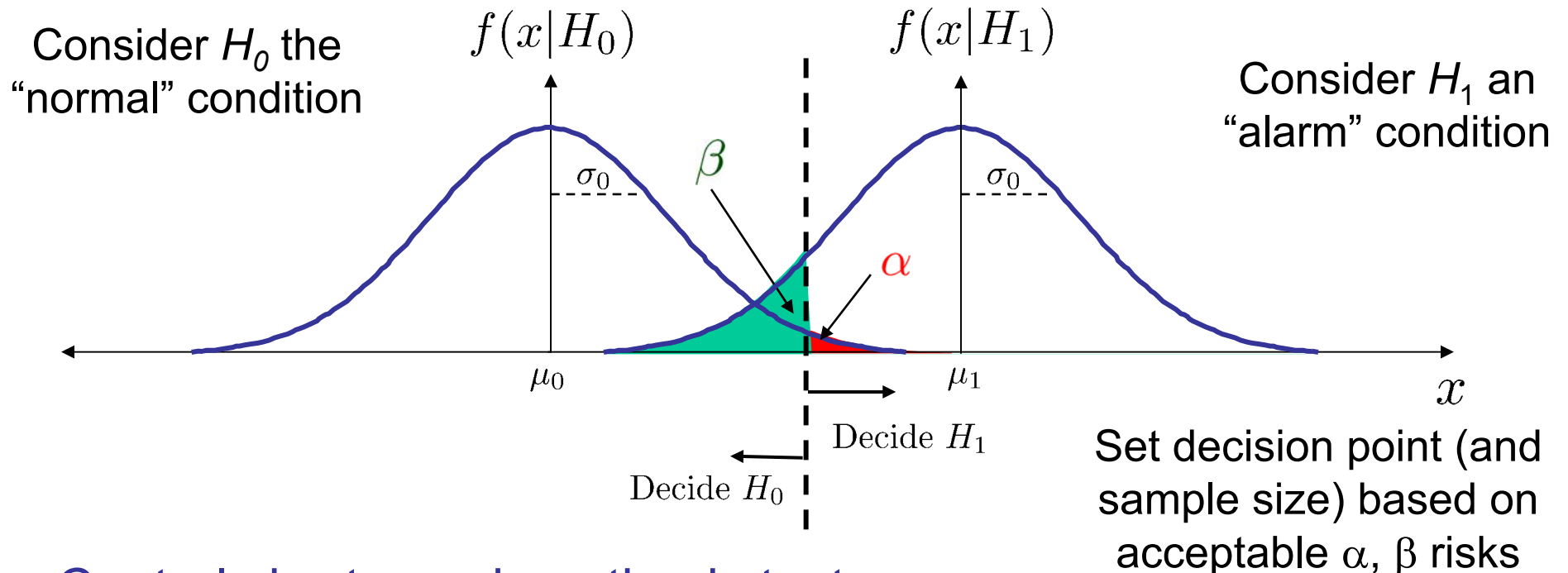
$$\alpha = \Pr(\text{reject } H_0 | H_0 \text{ is true})$$

- Type II error: “miss”

$$\beta = \Pr(\text{accept } H_0 | H_0 \text{ is false})$$

- Power of test (“correct alarm”)

$$1 - \beta = \Pr(\text{reject } H_0 | H_0 \text{ is false})$$



- Control charts are hypothesis tests:

- Is my process “in control” or has a significant change occurred?

Summary

1. Review: Probability Distributions & Random Variables
2. Sampling: Key distributions arising in sampling
 - Chi-square, t, and F distributions
3. Estimation: Reasoning about the population based on a sample
4. Some basic confidence intervals
 - Estimate of mean with variance known
 - Estimate of mean with variance not known
 - Estimate of variance
5. Hypothesis tests

Next Time:

1. Are effects (some variable) significant?
⇒ ANOVA (Analysis of Variance)
2. How do we model the effect of some variable(s)?
⇒ Regression modeling