Probability & Statistics for System Design

Doug Bodner

ASE 6002 – System Design and Analysis

Outcomes

- Understand basics of uncertainty
- Understand random variables
- Use probability distributions to model uncertain random variables
- Understand basic statistical methods

Uncertainty

- Uncertainty affects system performance
  - System performance in known usage can vary
  - System usage itself can vary
  - External factors affect system performance, and these often are uncertain
- How can we characterize uncertainty to help design robust systems?
- How do we imagine the possible?

Example

- The World Trade Center opened in 1973 as a state-of-the-art complex featuring the then-tallest buildings in the world
- On September 11, 2001, terrorists flew a jet into each twin tower, destroying them and killing thousands
- The designers of the building system never envisioned this possibility, nor did anyone else until the terrorists did
- Extreme example of external factors affecting the system
- "Unknown unknown" not accounted for in system design

By Michael Foran, CC BY 2.0, https://commons.wikimedia.org/w/index.php?curid=11785530

Other Examples

- Weight of passengers on motorcycle system (affecting top speed, mpg)

- Battery capacity tolerance (affecting battery life between recharges)

- Demand at e-commerce warehouse (affecting time to fill orders, labor requirements, etc.)

Types of Uncertainty

- Data-supported uncertainty
  - Often, we have data from past systems that help us understand the uncertainty in our to-be-designed system
  - Probability and statistics can be applied
  - Focus of these slides

- Uncertainty with little to no data
  - You will learn about this in the course

Random Variables

- Consider a system factor that experiences variability
  - Miles per gallon of a car
- Let X represent that variable factor, which we will call a *random variable*
- Take observations of that factor, each one x_i where i indexes the observations from 1 to some number n
- The data give us some insight into the random variable X
  - Mean and spread, for example

Visualizing Data with Histograms

- Histograms display frequency of observations x_i with different buckets or ranges of values
- Useful for large datasets
- Histograms display
  - Range
  - Frequency of observations in different buckets (i.e., estimate of likelihood)
- Rule of thumb – use approximately $\sqrt{n}$ buckets where n is the number of observations

Histograms (cont'd)

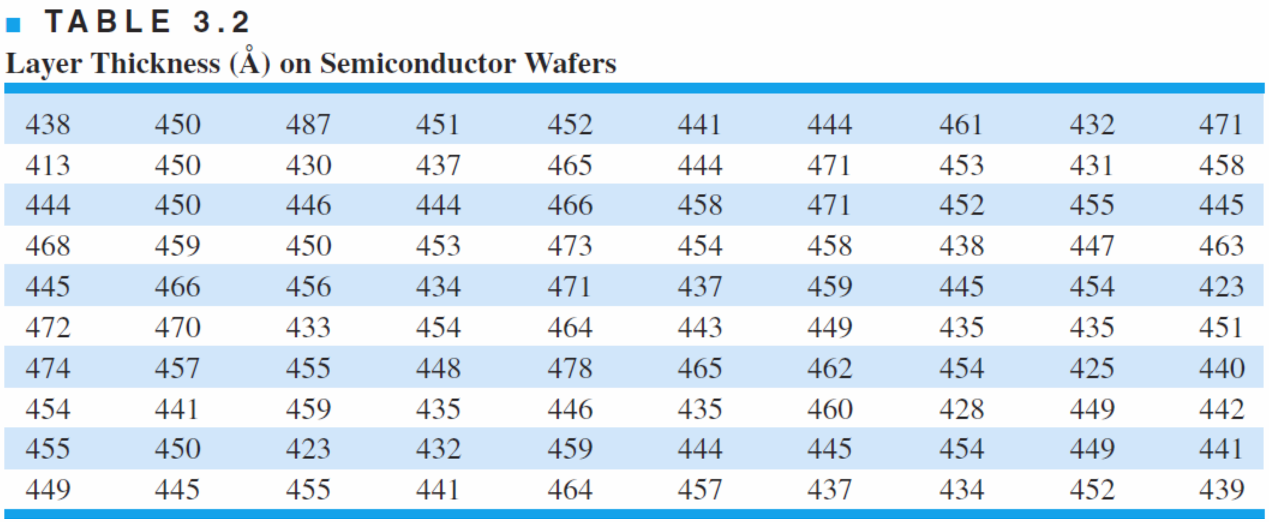


TABLE 3.2

Layer Thickness (Å) on Semiconductor Wafers

438	450	487	451	452	441	444	461	432	471
413	450	430	437	465	444	471	453	431	458
444	450	446	444	466	458	471	452	455	445
468	459	450	453	473	454	458	438	447	463
445	466	456	434	471	437	459	445	454	423
472	470	433	454	464	443	449	435	435	451
474	457	455	448	478	465	462	454	425	440
454	441	459	435	446	435	460	428	449	442
455	450	423	432	459	444	445	454	449	441
449	445	455	441	464	457	437	434	452	439

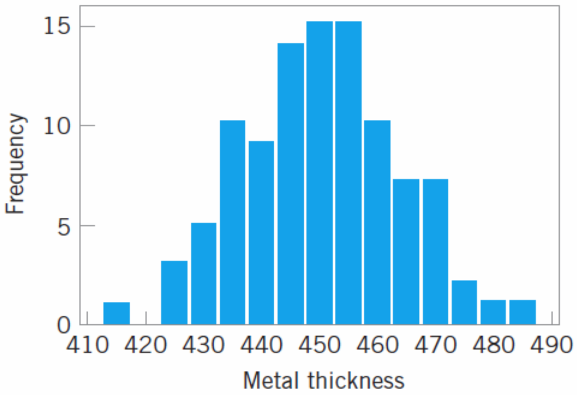


FIGURE 3.4 Minitab histogram with 15 bins for the metal layer thickness data.

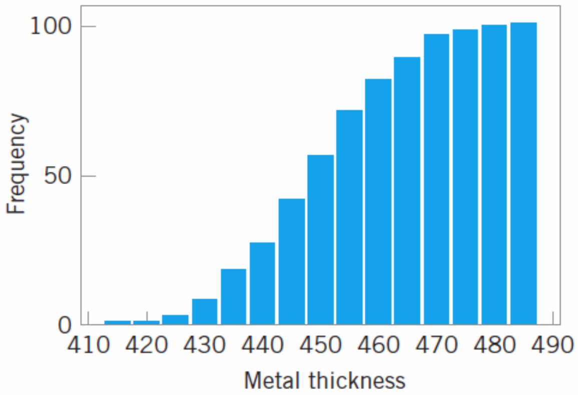
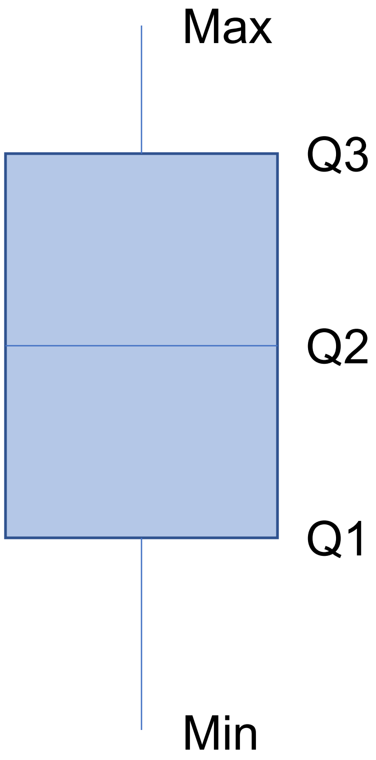


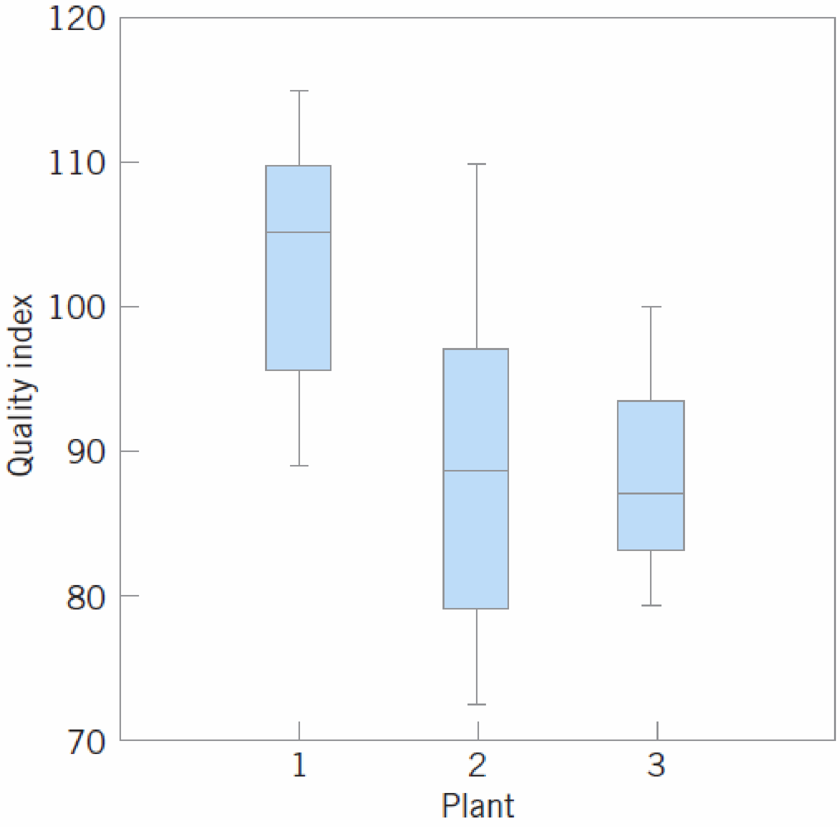
FIGURE 3.5 A cumulative frequency plot of the metal thickness data from Minitab.

Visualizing Data with Box Plots

- Visualizes several important aspects of a dataset:
  - Central tendency (median = Q2)
  - Variability (quartiles Q1 and Q3)
  - Symmetry or lack thereof (placement of median in quartile box)
  - Range (min and max)



Comparing Box Plots



■ **FIGURE 3.8** Comparative box plots of a quality index for products produced at three plants.

Are the plants producing at approximately the same quality level?

Sample Mean

- What is the central tendency of the data?
- Sample mean is generally considered the most important measure of central tendency
- Given n observations $x_1, x_2, \ldots, x_n$

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

Sample Median

- Sample mean is affected by lack of symmetry
  - One outlier can have a significant effect on the mean
- Sample median is considered more reliable in the sense of measuring central tendency without effect of outliers or significant lack of symmetry
- The median is the point at which the sample is divided into two equal halves

$$x_M = \begin{cases} x_{([n+1]/2)} & \text{if n odd} \\ \left[x_{(n/2)} + x_{([n/2]+1)}\right]/2 & \text{if n even} \end{cases}$$

Note: $x_{(i)}$ is an order statistic, the ith observation when observations are ordered by ascending value

Sample Variance

- Variability in sample data is measured by the sample variance
- Goal is to understand variance (or spread)
- Given n observations $x_1, x_2, \ldots, x_n$

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

Sample Standard Deviation

- The sample standard deviation expresses variability in the same units as the original data

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

Linear Combinations of Random Variables

- Let $Y = a_0 + a_1X_1 + a_2X_2 + a_3X_3 + \cdots + a_mX_m$
  - X_j are independently distributed random variables for $j = 1\ to\ m$ with mean μ_j and variance $\sigma_j{}^2$
  - a_j are constants for $j = 0\ to\ m$
- Population mean for Y
  - $\mu_Y = a_0 + a_1\mu_1 + a_2\mu_2 + a_3\mu_3 + \cdots + a_m\mu_m$
- Population variance for Y
  - $\sigma_Y{}^2 = a_1{}^2\sigma_1{}^2 + a_2{}^2\sigma_{22} + a_3{}^2\sigma_3{}^2 + \cdots + a_m{}^2\sigma_m{}^2$

Types of Random Variables

- If X is a random variable
- X can be discrete, meaning it takes on a limited number of values (sometimes just integers within an interval)
- X can be continuous, meaning it takes on real number values in an interval (infinite number of possibilities)

Probability Distributions

- We use probability distribution functions (PDFs) with mathematical forms to determine the probabilities of certain outcomes of the random variables for our design and analysis purposes
- When the RVs are discrete, these are called probability mass functions (PMFs)
- Many mathematical PMFs and PDFs exist
- We select a mathematical PMF or PDF based on the RV characteristics and the data observations

Probability Distributions (cont'd)

- Data is sampled from a population
- The population is all the possible observations of the system factor
  - All possible outcomes over time of that factor
  - Assume that the factor's parameters do not change
- Let X be a random variable representing the factor
- The probability distribution of X relates a particular value $X = x_i$ to the probability of x_i occurring in the population
- We use distributions to model certain occurrences

PMFs and PDFs

- For discrete RVs
  - For a value x_i in range R_X of X, probability that any $X = x_i$ is given by

$$p(x_i) = P\{X = x_i\}$$

- For continuous RVs
  - For any region r consisting of a set of values, each in R_X, the probability of a particular value of $X = x$ being in that range is the integral over r of

$$f(x)$$

- Note that $p(x_i)$ and $f(x)$ are characterized by parameters

Visualizing PMFs and PDFs

Probability mass function

Probability density function

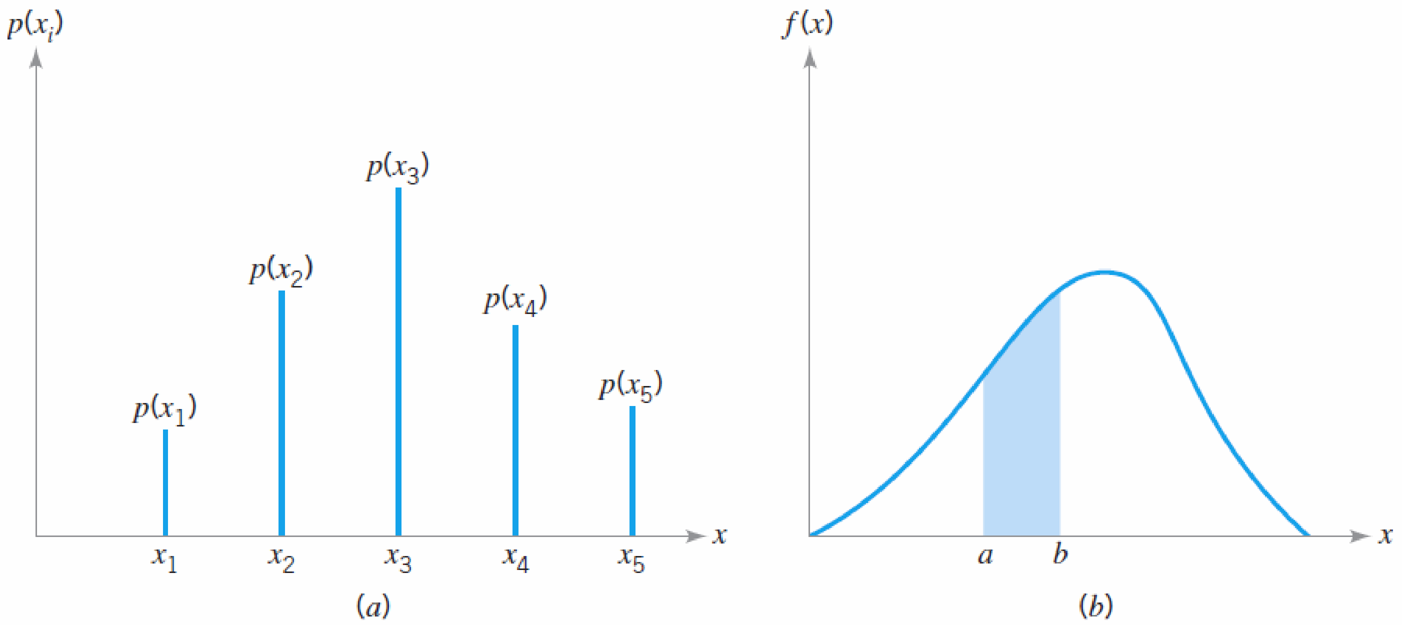
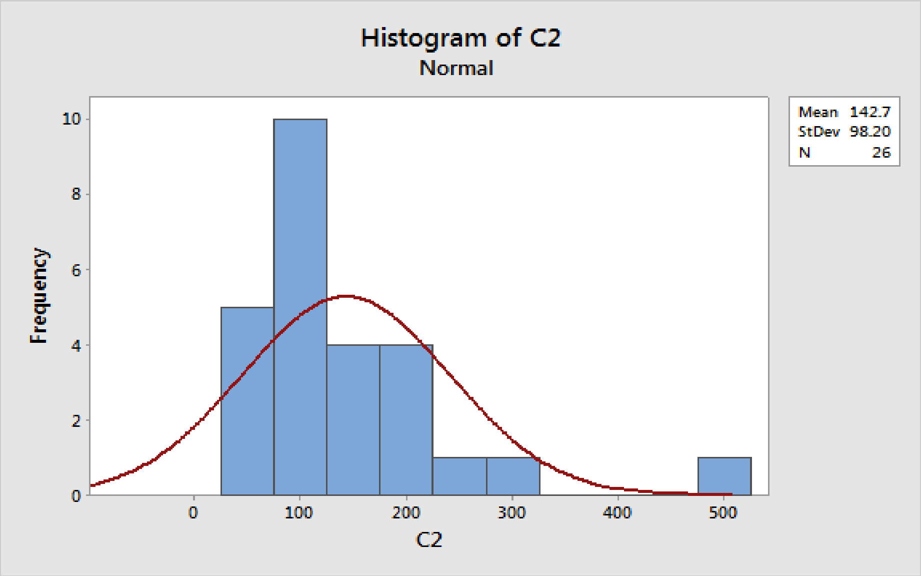


FIGURE 3.9 Probability distributions. (*a*) Discrete case. (*b*) Continuous case.

PDFs and Histograms

- The histogram of observations x_i of random variable X will tend to look like the pdf of X
- Aluminum contamination data with histogram and fitted PDF (red)
- Need more data for better fit

30	30	60	63
70	79	87	90
101	102	115	118
119	119	120	125
140	145	172	182
183	191	222	244
291	511		



CDFs

- The cumulative density or distribution function states the probability that $X \leq x$ for some value x

$$F(x) = \sum_{x_i \leq x} p(x_i) \ \text{for discrete } X$$

$$F(x) = \int_{-\infty}^{x} f(x) \ \text{for continuous } X$$

- $0 \leq F(x) \leq 1$

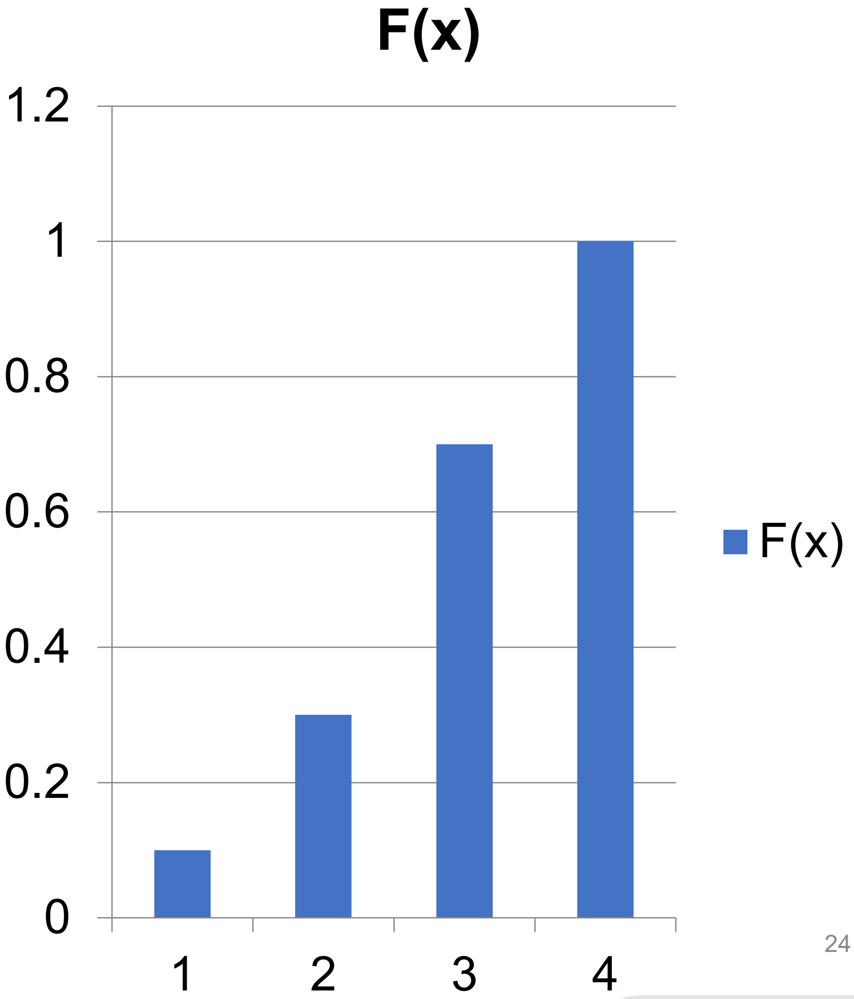
Visualizing CDFs

- Probability that a random variable $x \leq x_0$
- Discrete

$$F(x_0) = \sum_{x_i \leq x_0} p(x_i)$$

- Continuous

$$F(x_0) = \int_{-\infty}^{x_0} p(x) dx$$



Mean and Variance

- Mean μ is a measure of central tendency and is the average or expected value of X

$$\mu = E[X] = \sum_{\forall i} x_i p(x_i) \text{ or } \mu = \int_{-\infty}^{\infty} x f(x)$$

- Variance is a measure of dispersion

$$\sigma^2 = E[X^2] - \mu^2$$

- Coefficient of variation is ratio of stdev to mean σ/μ
- CV is important since it tells us how important variability is
- Small CV means variability may not be that important

Discrete Uniform Distribution

- Let n be the number of values of possible outcomes x_i

$$p(x_i) = \frac{1}{n}$$

- Assume integer values in the interval $[a, b]$ with no gaps

$$\mu = \frac{a+b}{2} \qquad \sigma^2 = \frac{n^2-1}{12}$$

- Parameters are $a =$ lower bound, $b =$ upper bound, $n =$ number of different values

Poisson Distribution

- Given a time interval or an area, how many events occur, given events are independent of one another
  - Customer arrivals per unit time
  - Defects per unit area
  - Number of crimes committed per unit time and area
- Parameter is λ = arrival rate > 0
- PMF is $p(x) = \frac{e^{-\lambda}\lambda^x}{x!}$

$$\mu = \lambda \quad \sigma^2 = \lambda$$

Poisson Additivity

- Sum of n Poisson random variables, each with parameter λ_i, is a Poisson RV with

$$\lambda = \sum_{i=1}^{n} \lambda_i$$

- What happens as n gets large?
- Via the Central Limit Theorem, the summed Poisson RV becomes approximately normally distributed

Continuous Uniform

- Equally probably outcomes for $x \in [a, b]$
- If $x \geq 0$, can set $a \geq 0$
- PDF is $f(x) = \begin{cases} \frac{1}{b-a} & \text{for } x \in [a, b] \\ 0 & \text{otherwise} \end{cases}$
- CDF is $F(x) = \begin{cases} 0 & \text{for } x < a \\ \frac{x-a}{b-a} & \text{for } x \in [a, b] \\ 1 & \text{for } x > b \end{cases}$

$$\mu = \frac{a+b}{2} \qquad \sigma^2 = \frac{(b-a)^2}{12}$$

Exponential Distribution

- Strongly related to Poisson distribution
- Can represent time between Poisson events
- Can represent time to an event (e.g., failure)
- Parameter is $\lambda > 0$
- Memoryless

$$P\{X > s + t | X > s\} = P\{X > t\}$$

Exponential PDF, CDF, Mean, Variance

- PDF

$$f(x) = \begin{cases} \lambda e^{-\lambda x} & \text{for } x \geq 0 \\ 0 & \text{otherwise} \end{cases}$$

- CDF

$$F(x) = \begin{cases} 0 & \text{for } x < 0 \\ 1 - e^{-\lambda x} & \text{for } \ x \geq 0 \end{cases}$$

- Mean and variance

$$\mu = \frac{1}{\lambda} \qquad \sigma^2 = \frac{1}{\lambda^2}$$

Normal Distribution

- Sum of different “component” RVs (via Central Limit Theorem)
- Parameters μ and σ
- Symmetric around mean
- Standard normal conversion $\Phi\left(\frac{x-\mu}{\sigma}\right)$
- Has unbounded tails, so may not be appropriate for factors that have upper or lower limits (e.g., only positive values)

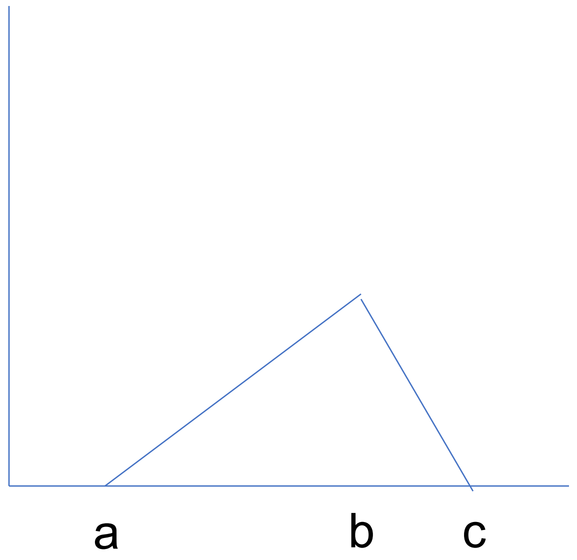
Triangular Distribution

- Commonly used in simulation
- Parameters

$$a = \text{lower bound}$$
$$b = \text{mode}$$
$$c = \text{upper bound}$$

- Useful in case of no data, only expert opinion

- PDF



Triangular PDF and CDF

- PDF

$$f(x) = \begin{cases} \dfrac{2(x-a)}{(b-a)(c-a)} & \text{for } a \le x \le b \\ \dfrac{2(c-x)}{(c-b)(c-a)} & \text{for } b < x \le c \\ 0 & \text{otherwise} \end{cases}$$

- CDF

$$F(x) = \begin{cases} 0 & \text{for } x \le a \\ \dfrac{(x-a)^2}{(b-a)(c-a)} & \text{for } a < x \le b \\ 1 - \dfrac{(c-x)^2}{(c-b)(c-a)} & \text{for } b < x \le c \\ 1 & \text{for } x > c \end{cases}$$

Triangular Mean and Variance

- $\mu = \frac{a+b+c}{3}$
- $\sigma^2 = \frac{a^2+b^2+c^2-ab-ac-bc}{18}$

PERT Distribution

- Continuous PDF with lower bound a, mode b and upper bound c
- Often used in project management for time durations
- Not as commonly used in simulation
- Based on beta distribution

$$\mu = \frac{a + 4b + c}{6} \qquad \sigma^2 = \frac{(\mu - a)(c - \mu)}{7}$$

Other Distributions

- Erlang
- Lognormal
- Weibull
- Beta
- Gamma
- Pearson

- More complex functional forms and parameter sets
- Typically need data to fit one of these

Common Distributions

Distribution	Applications
Uniform	Random variable when only lower and upper bounds known Equally likely values within an interval Task times Counts (discrete)
Triangular	Random variable when lower bound, upper bound and mode are known Task times
Poisson	Number of arrivals per time interval Number of defects per unit product
Normal	Population variables, sums of RVs (CLT) Measurements (e.g, length) Task times (composed of sum of many steps)
Exponential	Reliability / time to failure (constant failure rate over time) Interarrival times

Less Common Distributions

Distribution	Applications
Gamma	Task times Time to failure
Erlang	Sum of exponential variables Task times with multiple phases Failure times with redundant components
Weibull	Task times Time to failure
Lognormal	Task times Time to failure
Pert	Task times in bounded range with limited information
Beta	Task times in bounded range

Simulation Models and Distributions

- Model development is an iterative process
- Start simple with a prototype
- Insert simple, limited data distributions at first
  - Triangular
  - Uniform
  - Exponential and Poisson
- As model is refined, collect data and fit distributions for better accuracy

Input-Output

- We typically have many random variables as inputs into our simulation models

- We care about outputs (performance)

- These are complex functions of input random variables

- Outputs are also … random variables!

- How do we estimate their means and variance?

Estimation Terminology

- Estimator:
  - Any function of a random sample, which is used to estimate population parameters
- Point estimate:
  - A single numerical value as the estimate of the unknown parameter
- Interval estimate:
  - A random interval (or called confidence interval) in which the true value of the parameter falls with some level of probability

Point Estimation

- Estimators are themselves random variables
  - $\hat{\mu} = (x_1 + x_2 + \cdots + x_n)/n = \bar{x}$
  - $\widehat{\sigma^2} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 = s^2$
    - $\hat{\sigma} = s/c_4$ (best)
    - $\hat{\sigma} = R/d_2$ (easy to calculate)

- Desired properties
  - Unbiased – the expected value of the estimate should equal the parameter being estimated
  - Minimum variance – the estimator should have the smallest variance among all possible estimators

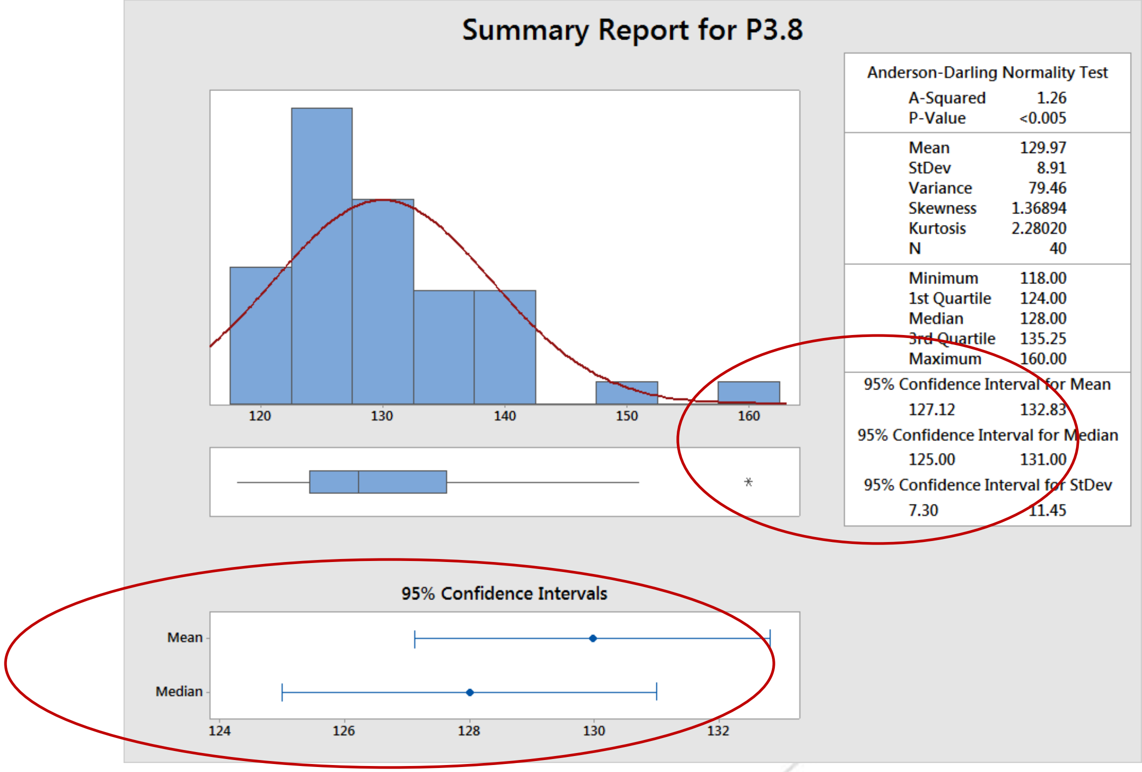
Estimators for Common Distributions

Distribution	Parameter	Estimator
Normal	μ	$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i = \bar{x}$
	σ^2	$\widehat{\sigma^2} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 = s^2$
	σ	It turns out that s is not an unbiased estimator for σ. Two alternatives: • $\hat{\sigma} = s/c_4$ (best) • $\hat{\sigma} = R/d_2$ (easy to calculate)
Poisson	λ	$\hat{\lambda} = \frac{1}{n}\sum_{i=1}^{n} x_i = \bar{x}$
Binomial	p	$\hat{p} = \frac{1}{n}\sum_{i=1}^{n} x_i = \bar{x}$, where $x_i \in \{0,1\}\ \forall\ i$

Note: $R = \text{range} = \max_i(x_i) - \min_i(x_i)$. Values of c_4 and d_2 are in Appendix VI in the book for various values of n

Confidence Interval Concepts

- What is an interval that contains the true value of a parameter with some confidence $100(1-\alpha)\%$?
  - Upper limit
  - Lower limit
  - Both these are statistics



Summary Report for P3.8

Anderson-Darling Normality Test

A-Squared	1.26
P-Value	<0.005
Mean	129.97
StDev	8.91
Variance	79.46
Skewness	1.36894
Kurtosis	2.28020
N	40
Minimum	118.00
1st Quartile	124.00
Median	128.00
3rd Quartile	135.25
Maximum	160.00

95% Confidence Interval for Mean

127.12 132.83

95% Confidence Interval for Median

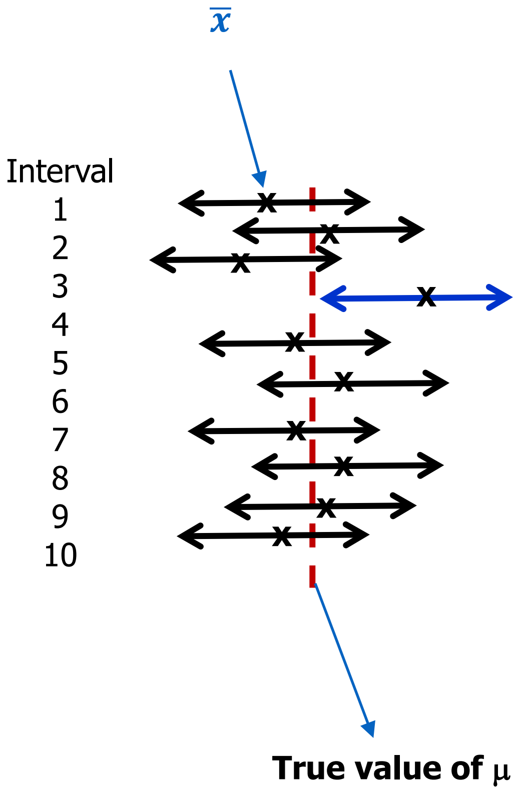
125.00 131.00

95% Confidence Interval for StDev

7.30 11.45

Confidence Interval Interpretation

- Calculate a 95% CI for μ based on a sample of 4 observations.
- Collect 100 samples and compute sample means for each
- Compute 100 CIs
  - (0.85,1.15), (0.8, 1.1), (0.9,1.0),…
  - Some of those computed CIs may contain the true mean, and some may not contain the true mean
- "95% confidence" means that in the long run, 95% of all computed CIs will contain the true mean μ.
  - In other words, 5% of these computed CIs will not trap the true mean



Confidence Interval Estimation

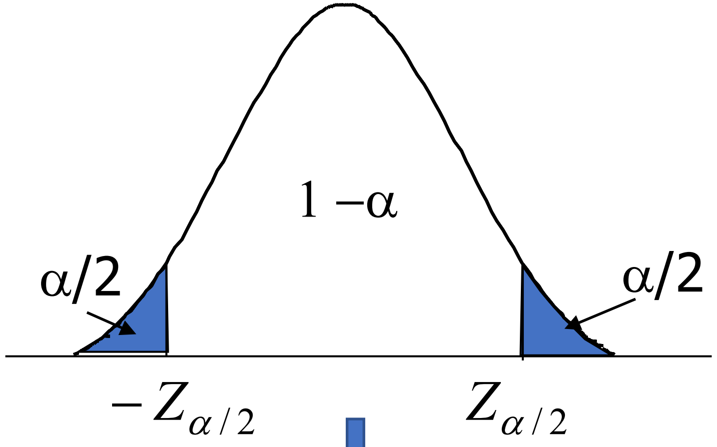
- Mean
  - Normally distributed data
    - Known $\sigma \rightarrow$ use Normal distribution
    - Unknown $\sigma \rightarrow$ use t distribution
  - Non-normal data with large sample size ($n > 30$)
    - Known/unknown $\sigma \rightarrow$ use Normal distribution
    - (based on central limit theorem)
- Variance
  - Normally distributed data → use Chi-Square distribution
  - Won’t cover in this class

CI for Mean of Normal Population

- Two-sided (1 - ∝)*100% confidence interval
- Mean μ of normal population
- Variance σ^2 known

- $x_1, x_2, \cdots x_n \sim NID(\mu, \sigma^2)$
- $\bar{x} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$ by CLT
- $\frac{\bar{x}-\mu}{\sigma/\sqrt{n}} \sim N(0, 1)$ by transform

- Note: This CI can be used for mean of non-normal distributions when $n > 30$



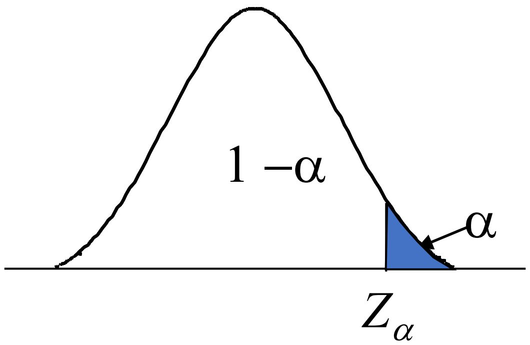
$$\boxed{P\left\{-z_{\alpha/2} \le \frac{\bar{x}-\mu}{\sigma/\sqrt{n}} \le z_{\alpha/2}\right\} = 1-\alpha}$$

$$\boxed{\bar{x} - Z_{\alpha/2} \times \frac{\sigma}{\sqrt{n}}} \le \mu \le \boxed{\bar{x} + Z_{\alpha/2} \times \frac{\sigma}{\sqrt{n}}}$$

CI Lower **CI Upper**

CI for Mean of Normal Population

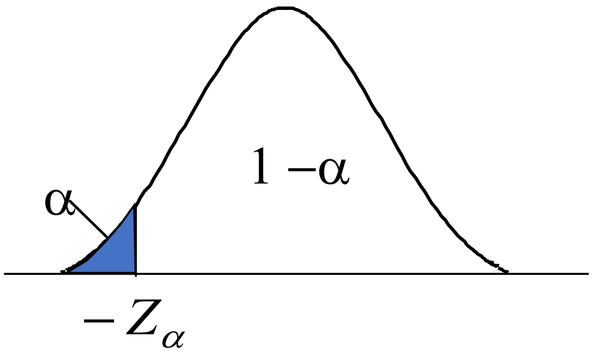
- One-sided (1 - ∝)*100% confidence interval
- Mean μ of normal population with variance σ^2 known



$$\Pr\left\{\frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \le Z_\alpha\right\} = 1-\alpha$$

$$\boxed{\mu \ge \bar{X} - Z_\alpha \times \frac{\sigma}{\sqrt{n}}}$$

CI Lower



$$\Pr\left\{-Z_\alpha \le \frac{\bar{X}-\mu}{\sigma/\sqrt{n}}\right\} = 1-\alpha$$

$$\boxed{\mu \le \bar{X} + Z_\alpha \times \frac{\sigma}{\sqrt{n}}}$$

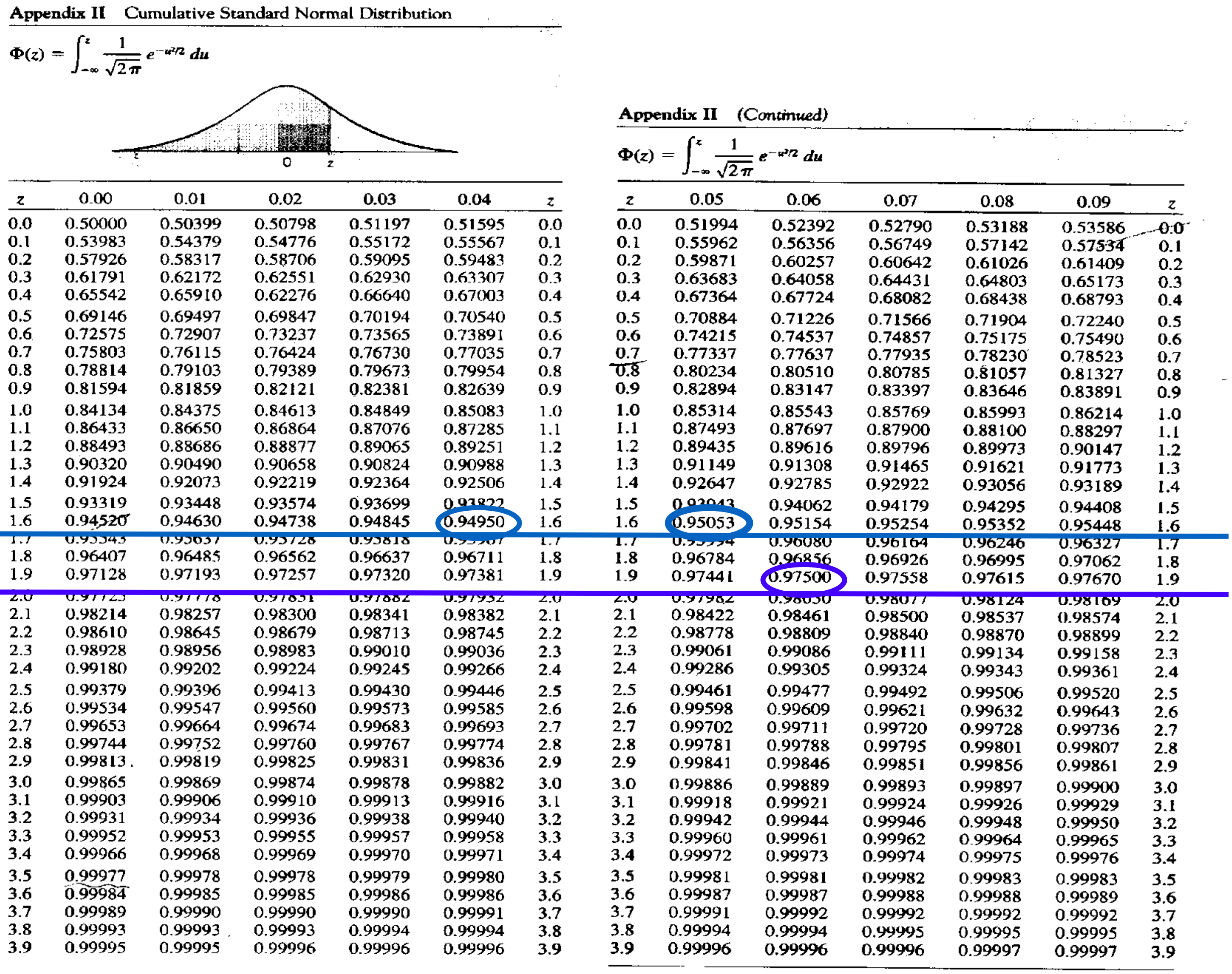
CI Upper

Note: This CI can be used for mean of non-normal distributions when n > 30

CI with Known σ Example

- The strength of a disposable plastic beverage container is being investigated. The strengths are normally distributed, with a known standard deviation of 15 psi. A sample of 20 plastic containers has a mean strength of 246 psi. Compute the 95% two-sided CI for the process mean.
  - $\bar{x} = 246, \sigma = 15, n = 20, \alpha = 0.05$
  - $Z_{\alpha/2} = \Phi^{-1}(0.975) = 1.96$
  - 95% Confidence interval for μ
    - $\bar{x} - Z_{a/2}\frac{\sigma}{\sqrt{n}} \le \mu \le \bar{x} + Z_{a/2}\frac{\sigma}{\sqrt{n}}$
    - $\bar{x} \pm Z_{a/2}\frac{\sigma}{\sqrt{n}} = 246 \pm 1.96\frac{15}{\sqrt{20}} = 246 \pm 6.57$

Normal Table Look-up



Appendix II Cumulative Standard Normal Distribution

$$\Phi(z) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} e^{-u^2/2}\, du$$

Appendix II *(Continued)*

$$\Phi(z) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} e^{-u^2/2}\, du$$

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09	z
0.0	0.50000	0.50399	0.50798	0.51197	0.51595	0.51994	0.52392	0.52790	0.53188	0.53586	0.0
0.1	0.53983	0.54379	0.54776	0.55172	0.55567	0.55962	0.56356	0.56749	0.57142	0.57534	0.1
0.2	0.57926	0.58317	0.58706	0.59095	0.59483	0.59871	0.60257	0.60642	0.61026	0.61409	0.2
0.3	0.61791	0.62172	0.62551	0.62930	0.63307	0.63683	0.64058	0.64431	0.64803	0.65173	0.3
0.4	0.65542	0.65910	0.62276	0.66640	0.67003	0.67364	0.67724	0.68082	0.68438	0.68793	0.4
0.5	0.69146	0.69497	0.69847	0.70194	0.70540	0.70884	0.71226	0.71566	0.71904	0.72240	0.5
0.6	0.72575	0.72907	0.73237	0.73565	0.73891	0.74215	0.74537	0.74857	0.75175	0.75490	0.6
0.7	0.75803	0.76115	0.76424	0.76730	0.77035	0.77337	0.77637	0.77935	0.78230	0.78523	0.7
0.8	0.78814	0.79103	0.79389	0.79673	0.79954	0.80234	0.80510	0.80785	0.81057	0.81327	0.8
0.9	0.81594	0.81859	0.82121	0.82381	0.82639	0.82894	0.83147	0.83397	0.83646	0.83891	0.9
1.0	0.84134	0.84375	0.84613	0.84849	0.85083	0.85314	0.85543	0.85769	0.85993	0.86214	1.0
1.1	0.86433	0.86650	0.86864	0.87076	0.87285	0.87493	0.87697	0.87900	0.88100	0.88297	1.1
1.2	0.88493	0.88686	0.88877	0.89065	0.89251	0.89435	0.89616	0.89796	0.89973	0.90147	1.2
1.3	0.90320	0.90490	0.90658	0.90824	0.90988	0.91149	0.91308	0.91465	0.91621	0.91773	1.3
1.4	0.91924	0.92073	0.92219	0.92364	0.92506	0.92647	0.92785	0.92922	0.93056	0.93189	1.4
1.5	0.93319	0.93448	0.93574	0.93699	0.93822	0.93943	0.94062	0.94179	0.94295	0.94408	1.5
1.6	0.94520	0.94630	0.94738	0.94845	0.94950	0.95053	0.95154	0.95254	0.95352	0.95448	1.6
1.7	0.95543	0.95637	0.95728	0.95818	0.95907	0.95994	0.96080	0.96164	0.96246	0.96327	1.7
1.8	0.96407	0.96485	0.96562	0.96637	0.96711	0.96784	0.96856	0.96926	0.96995	0.97062	1.8
1.9	0.97128	0.97193	0.97257	0.97320	0.97381	0.97441	0.97500	0.97558	0.97615	0.97670	1.9
2.0	0.97725	0.97778	0.97831	0.97882	0.97932	0.97982	0.98030	0.98077	0.98124	0.98169	2.0
2.1	0.98214	0.98257	0.98300	0.98341	0.98382	0.98422	0.98461	0.98500	0.98537	0.98574	2.1
2.2	0.98610	0.98645	0.98679	0.98713	0.98745	0.98778	0.98809	0.98840	0.98870	0.98899	2.2
2.3	0.98928	0.98956	0.98983	0.99010	0.99036	0.99061	0.99086	0.99111	0.99134	0.99158	2.3
2.4	0.99180	0.99202	0.99224	0.99245	0.99266	0.99286	0.99305	0.99324	0.99343	0.99361	2.4
2.5	0.99379	0.99396	0.99413	0.99430	0.99446	0.99461	0.99477	0.99492	0.99506	0.99520	2.5
2.6	0.99534	0.99547	0.99560	0.99573	0.99585	0.99598	0.99609	0.99621	0.99632	0.99643	2.6
2.7	0.99653	0.99664	0.99674	0.99683	0.99693	0.99702	0.99711	0.99720	0.99728	0.99736	2.7
2.8	0.99744	0.99752	0.99760	0.99767	0.99774	0.99781	0.99788	0.99795	0.99801	0.99807	2.8
2.9	0.99813	0.99819	0.99825	0.99831	0.99836	0.99841	0.99846	0.99851	0.99856	0.99861	2.9
3.0	0.99865	0.99869	0.99874	0.99878	0.99882	0.99886	0.99889	0.99893	0.99897	0.99900	3.0
3.1	0.99903	0.99906	0.99910	0.99913	0.99916	0.99918	0.99921	0.99924	0.99926	0.99929	3.1
3.2	0.99931	0.99934	0.99936	0.99938	0.99940	0.99942	0.99944	0.99946	0.99948	0.99950	3.2
3.3	0.99952	0.99953	0.99955	0.99957	0.99958	0.99960	0.99961	0.99962	0.99964	0.99965	3.3
3.4	0.99966	0.99968	0.99969	0.99970	0.99971	0.99972	0.99973	0.99974	0.99975	0.99976	3.4
3.5	0.99977	0.99978	0.99978	0.99979	0.99980	0.99981	0.99981	0.99982	0.99983	0.99983	3.5
3.6	0.99984	0.99985	0.99985	0.99986	0.99986	0.99987	0.99987	0.99988	0.99988	0.99989	3.6
3.7	0.99989	0.99990	0.99990	0.99990	0.99991	0.99991	0.99992	0.99992	0.99992	0.99992	3.7
3.8	0.99993	0.99993	0.99993	0.99994	0.99994	0.99994	0.99994	0.99995	0.99995	0.99995	3.8
3.9	0.99995	0.99995	0.99996	0.99996	0.99996	0.99996	0.99996	0.99996	0.99997	0.99997	3.9

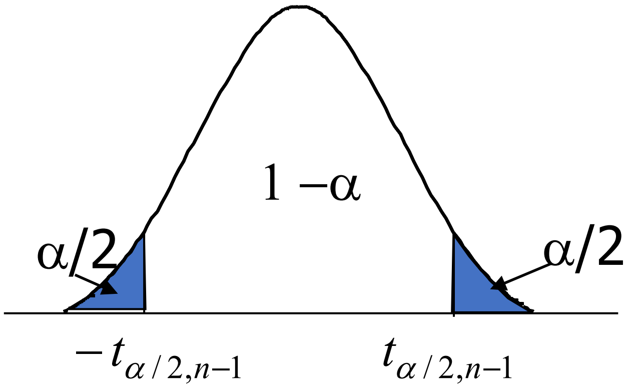
CI for Mean of Normal Population

- Two-sided (1 - ∝)*100% confidence interval
- Mean μ of normal population
- Variance σ^2 unknown

- $x_1, x_2, \cdots x_n \sim NID(\mu, \sigma^2)$
- $\frac{\bar{x}-\mu}{s/\sqrt{n}} \sim t(n-1)$

- Note:

$$t_{\alpha/2,n-1} \approx Z_{\alpha/2}; \text{for } n > 30$$

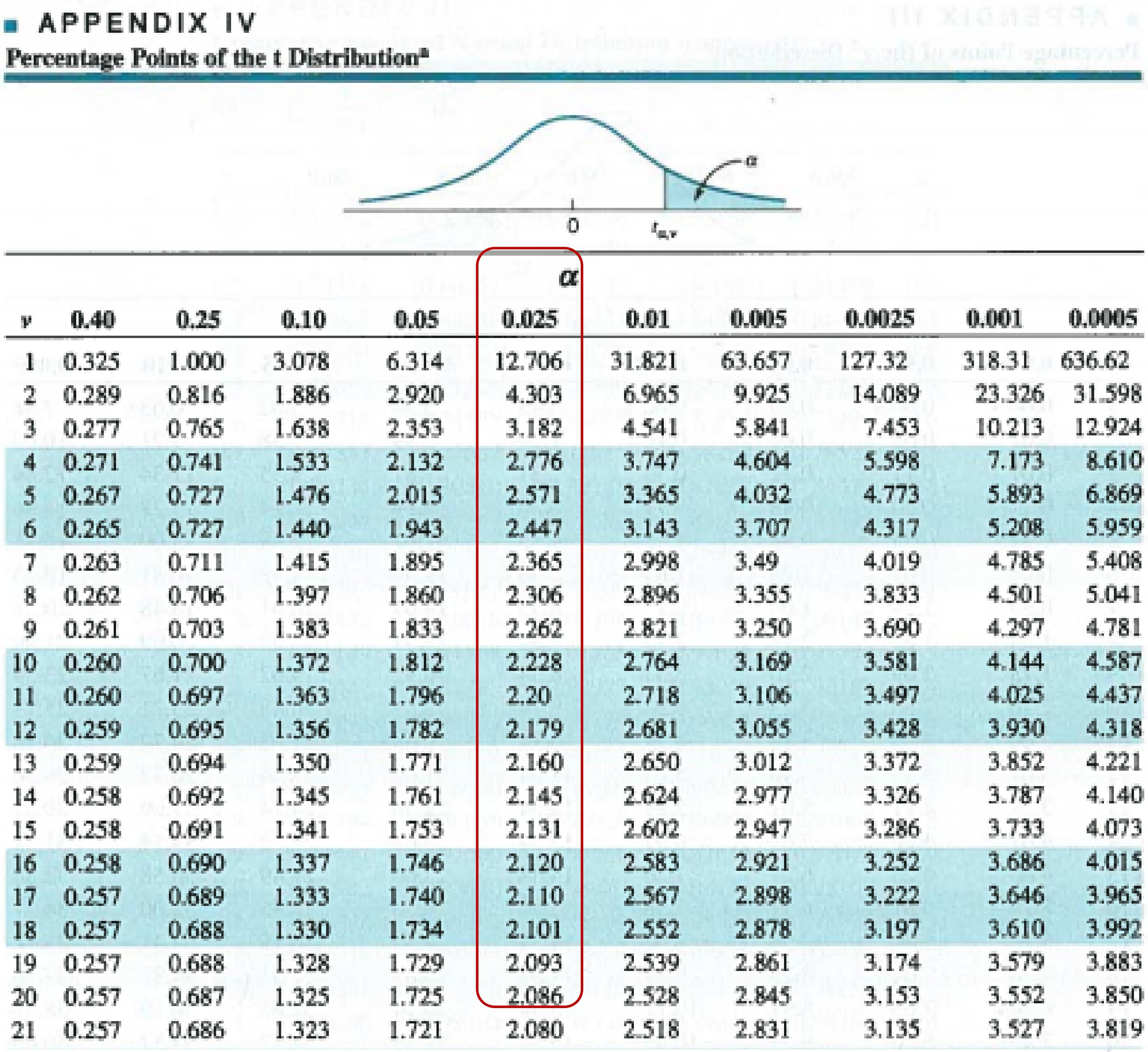


$$P\left\{-t_{\alpha/2,n-1} \le \frac{\bar{x}-\mu}{s/\sqrt{n}} \le t_{\alpha/2,n-1}\right\} = 1-\alpha$$

$$\boxed{\bar{x} - t_{\alpha/2,n-1}s/\sqrt{n}} \le \mu \le \boxed{\bar{x} + t_{\alpha/2,n-1}s/\sqrt{n}}$$

CI Lower **CI Upper**

t Table Look-up


α
0
$t_{\alpha,\nu}$

■ APPENDIX IV

Percentage Points of the t Distribution[a]

					α					
ν	**0.40**	**0.25**	**0.10**	**0.05**	**0.025**	**0.01**	**0.005**	**0.0025**	**0.001**	**0.0005**
1	0.325	1.000	3.078	6.314	12.706	31.821	63.657	127.32	318.31	636.62
2	0.289	0.816	1.886	2.920	4.303	6.965	9.925	14.089	23.326	31.598
3	0.277	0.765	1.638	2.353	3.182	4.541	5.841	7.453	10.213	12.924
4	0.271	0.741	1.533	2.132	2.776	3.747	4.604	5.598	7.173	8.610
5	0.267	0.727	1.476	2.015	2.571	3.365	4.032	4.773	5.893	6.869
6	0.265	0.727	1.440	1.943	2.447	3.143	3.707	4.317	5.208	5.959
7	0.263	0.711	1.415	1.895	2.365	2.998	3.49	4.019	4.785	5.408
8	0.262	0.706	1.397	1.860	2.306	2.896	3.355	3.833	4.501	5.041
9	0.261	0.703	1.383	1.833	2.262	2.821	3.250	3.690	4.297	4.781
10	0.260	0.700	1.372	1.812	2.228	2.764	3.169	3.581	4.144	4.587
11	0.260	0.697	1.363	1.796	2.20	2.718	3.106	3.497	4.025	4.437
12	0.259	0.695	1.356	1.782	2.179	2.681	3.055	3.428	3.930	4.318
13	0.259	0.694	1.350	1.771	2.160	2.650	3.012	3.372	3.852	4.221
14	0.258	0.692	1.345	1.761	2.145	2.624	2.977	3.326	3.787	4.140
15	0.258	0.691	1.341	1.753	2.131	2.602	2.947	3.286	3.733	4.073
16	0.258	0.690	1.337	1.746	2.120	2.583	2.921	3.252	3.686	4.015
17	0.257	0.689	1.333	1.740	2.110	2.567	2.898	3.222	3.646	3.965
18	0.257	0.688	1.330	1.734	2.101	2.552	2.878	3.197	3.610	3.992
19	0.257	0.688	1.328	1.729	2.093	2.539	2.861	3.174	3.579	3.883
20	0.257	0.687	1.325	1.725	2.086	2.528	2.845	3.153	3.552	3.850
21	0.257	0.686	1.323	1.721	2.080	2.518	2.831	3.135	3.527	3.819

CI for Mean of Normal Population

- One-sided (1 - ∝)*100% confidence interval
- Mean μ of normal population
- Variance σ^2 unknown
- $\bar{x}$ and s are based on sample data from the normal population
- One-sided upper confidence limit
  - $\mu \leq \bar{x} + t_{\alpha,n-1}\frac{s}{\sqrt{n}}$
- Lower confidence limit
  - $\mu \geq \bar{x} - t_{\alpha,n-1}\frac{s}{\sqrt{n}}$

CI with Unknown σ Example

- Suppose the mean load at failure from a tensile adhesion test is 13.71 Mpa, with $s = 3.55$ and $n = 21$.
  - What is a two-sided 95% confidence interval for the mean?
    - $\bar{x} \pm t_{\alpha/2,n-1} \frac{s}{\sqrt{n}} = 13.71 \pm t_{0.025,21-1} \frac{3.55}{\sqrt{21}} = 13.71 \pm 2.086(0.775) = (12.09, 15.33)$
  - What is a one-sided 95% confidence interval lower limit for the mean?
    - $\bar{x} - t_{\alpha,n-1} \frac{s}{\sqrt{n}} = 13.71 - t_{0.05,21-1} \frac{3.55}{\sqrt{21}} = 13.71 - 1.725(0.775) = 12.37$

Hypothesis Testing

- Drawing a conclusion based on statistical inference
- Closely related to confidence intervals
- Designing the hypothesis and test
  - State null hypothesis H_0
  - Frame alternate hypothesis H_1 relative to goal of study (one-sided versus two-sided)
  - Determine sample size n and level of significance α
  - Determine test statistic
  - Find the distribution of the test statistic and the rejection region of H_0

- Example: design & perform a test on the mean of a normal distribution with $\sigma^2 = 4$

$$H_0 : \mu = 1.5$$
$$H_1 : \mu \neq 1.5$$
$$n = 5$$
$$\alpha = 0.05$$

$$Z_0 = \frac{\overline{X} - \mu}{\sigma / \sqrt{n}}$$

Test Statistic

$$Z_0 \sim N(0,1)$$
$$|Z_0| > Z_{\alpha/2} = Z_{0.025} = 1.96$$

Rejection Region

Hypothesis Testing (cont'd)

- Once the test is designed, perform it
  - Collect sample data and calculate the test statistic using the sample
  - Compare the test statistic of the sample with the rejection region
  - Make the decision and assess the risk
- Example (cont'd)

$$\begin{aligned} X_1 &= 0.5 \\ X_2 &= 1.5 \\ X_3 &= 1 \\ X_4 &= 1.5 \\ X_5 &= 2 \end{aligned} \rightarrow \overline{X} = 1.3 \rightarrow Z_0 = \frac{1.3 - 1.5}{2 / \sqrt{5}} = -0.22 \qquad |Z_0| \ngtr Z_{\alpha/2} = 1.96$$

H_0 is not rejected

Hypothesis Testing Analogy

- Hypothesis testing in science is similar to the criminal court system in the United States. How do we decide guilt?
  - Assume innocence (H_0) until "proven" guilty (reject H_0)
  - Evidence is presented at a trial (sample)
  - Proof has to be "beyond a reasonable doubt" (statistical inference)
  - Show strong evidence to "prove" guilty (reject H_0)

- A jury's possible decision:
  - Guilty (reject H_0)
  - Not guilty (cannot reject H_0)

- Note that a jury cannot declare somebody "innocent"
- A jury can only say "not guilty"

- What are the risks?

Type I and Type II Errors

- Reject H_0 when it is true (convict when innocent)
  - Type I error
  - $\alpha = P\{\text{type I error}\} = P\{\text{reject } H_0 | H_0 \text{ is true}\}$
- Fail to reject H_0 when it is false (find not guilty when guilty)
  - Type II error
  - $\beta = P\{\text{type II error}\} = P\{\text{do not reject } H_0 | H_0 \text{ is false}\}$
- Power of test = $1 - \beta = P\{\text{reject } H_0 | H_0 \text{ is false}\}$

Methods for Hypothesis Testing

- Rejection region
  - Compute test statistic and cut-off value
  - Reject H_o if test statistic beyond cut-off
- Confidence interval
  - Compute test statistic and confidence interval
  - Reject H_o if hypothesized parameter value not in CI
- p-value
  - Compute probability of test statistic value given H_o
  - Reject H_o if this probability is too small

Typical Tests

- Two means are equal versus not equal
  - Comparing performance of two system designs
- One mean is better than another
  - Once again comparing performance of two system designs

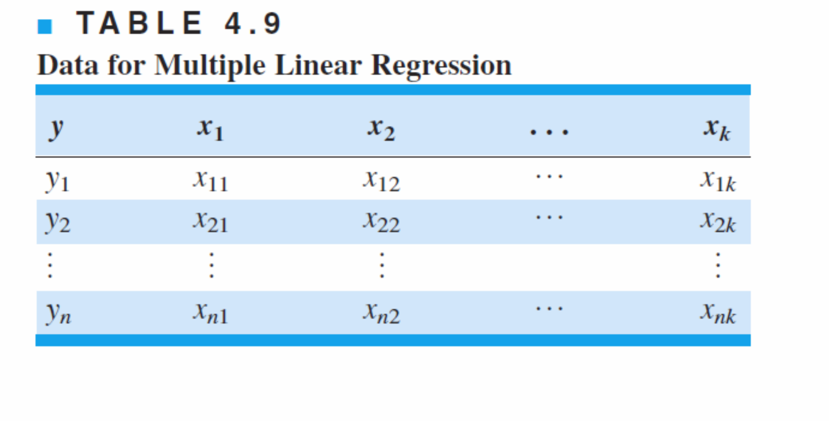
- There are statistical methods to do this, but we will not address them in the class
- Rather, we will find the optimal performance over many possible system designs!

Linear Regression

- Can a performance attribute be posed as a function of different predictor variables
- $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots \beta_k x_k + \varepsilon$
- Components
  - y is the response variable (dependent variable)
  - x_j are the predictor variables (independent variables)
  - β_j are the regression coefficients (constant parameters)
  - β_0 is the intercept term
  - $\varepsilon \sim N(0, \sigma^2)$ is the random error component
- Note: The regression equation is linear in β_j, so x_j can be polynomial functions
- Use regression to predict output (performance values) based on inputs (controllable factors or design variables)

Regression Data Sampling

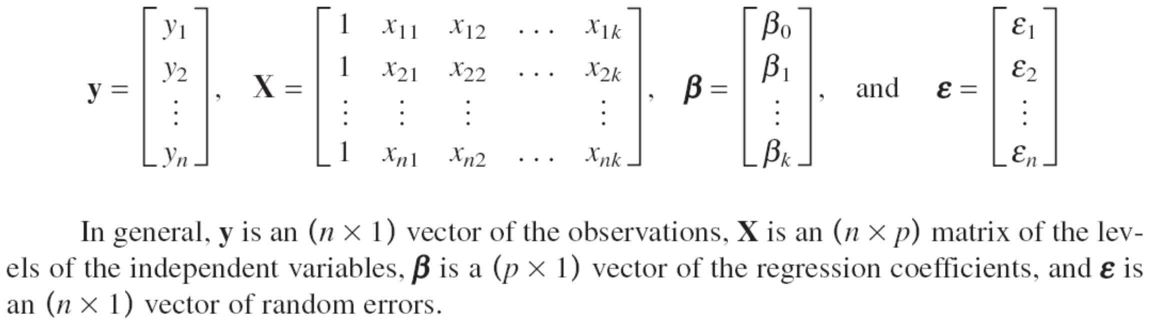
- i = 1, 2, …, n samples
- Each sample has one response variable and k predictor variables
- Resulting equations
  - $y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots \beta_k x_{ik} + \varepsilon_i$
  - $y_i = \beta_0 + \sum_{j=1}^{k} \beta_j x_{ij} + \varepsilon_i$
  - $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$



■ TABLE 4.9

Data for Multiple Linear Regression

y	x_1	x_2	$\dots$	x_k
y_1	x_{11}	x_{12}	$\cdots$	x_{1k}
y_2	x_{21}	x_{22}	$\cdots$	x_{2k}
$\vdots$	$\vdots$	$\vdots$		$\vdots$
y_n	x_{n1}	x_{n2}	$\cdots$	x_{nk}



$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, \quad \mathbf{X} = \begin{bmatrix} 1 & x_{11} & x_{12} & \dots & x_{1k} \\ 1 & x_{21} & x_{22} & \dots & x_{2k} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_{n1} & x_{n2} & \dots & x_{nk} \end{bmatrix}, \quad \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix}, \quad \text{and} \quad \boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

In general, $\mathbf{y}$ is an $(n \times 1)$ vector of the observations, $\mathbf{X}$ is an $(n \times p)$ matrix of the levels of the independent variables, $\boldsymbol{\beta}$ is a $(p \times 1)$ vector of the regression coefficients, and $\boldsymbol{\varepsilon}$ is an $(n \times 1)$ vector of random errors.

Method of Least Squares

- Minimize the sum of errors squared
  - Min $\sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{k} \beta_j x_j\right)^2$
  - Solve for β’s
- Solving these equations yields
  - $\mathbf{X}'\mathbf{X}\widehat{\boldsymbol{\beta}} = \mathbf{X}'\mathbf{y}$
  - $\widehat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$
- Fitted regression model
  - $\hat{\mathbf{y}} = \mathbf{X}\widehat{\boldsymbol{\beta}}$

Residuals

- Residuals
  - $\mathbf{e} = \mathbf{y} - \hat{\mathbf{y}}$
- Sum of squares of residuals
  - $SS_E = \mathbf{y}'\mathbf{y} - \widehat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{y}$
- Estimating variance
  - $E(SS_E) = \sigma^2(n-p)$ where $p = k + 1$
  - $\hat{\sigma}^2 = \frac{E(SS_E)}{(n-p)}$ is an unbiased estimator

Hypothesis Testing

- Sum of squares
  - $SS_T = SS_R + SS_E$
- Hypotheses
  - $H_0\colon \beta_j = 0\ \forall j$
  - $H_1\colon \beta_j \neq 0$ for some j
- Test statistic $F_0 = \dfrac{SS_R/k}{SS_E/(n-k-1)} = \dfrac{MS_R}{MS_E} \sim F_{k,n-k-1}$
- Reject H_0 if $F_0 > F_{\alpha,k,n-k-1}$
  - Reject H_0 if the variation due to regression is substantially greater than that due to randomness

Other Measures of Model Fit

- Percent reduction in variance from regression coefficients
  - $R^2 = \frac{SS_R}{SS_T} = 1 - \frac{SS_E}{SS_T}$
  - How well does the model explain the variation
- Adjusted by number of regression terms
  - $R_{adj}^2 = 1 - \frac{SS_E/(n-p)}{SS_t/(n-1)} = 1 - \left(\frac{n-1}{n-p}\right)(1 - R^2)$
  - With enough terms, $R^2 \to 1$
  - R_{adj}^2 compensates for the number of terms
  - If $R_{adj}^2 \ll R^2$, the model has too many insignificant terms

Regression Example

- The brake horsepower developed by an automobile engine on a dynamometer is thought to be a function of:
  - the engine speed in revolutions per minute (rpm)
  - the road octane number of the fuel
  - the engine compression.
- An experiment is run in the laboratory

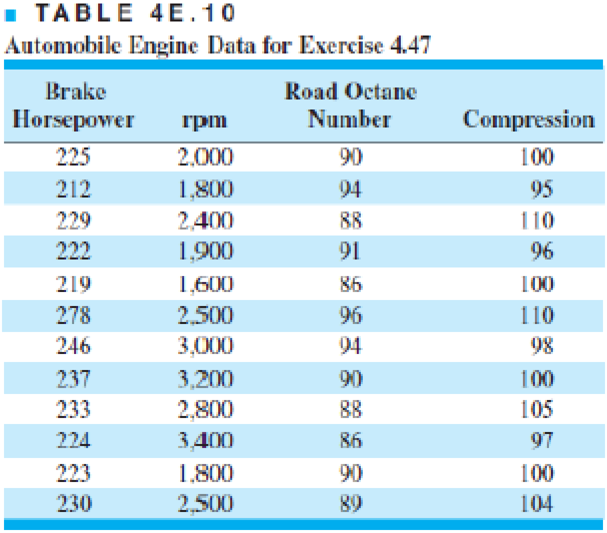


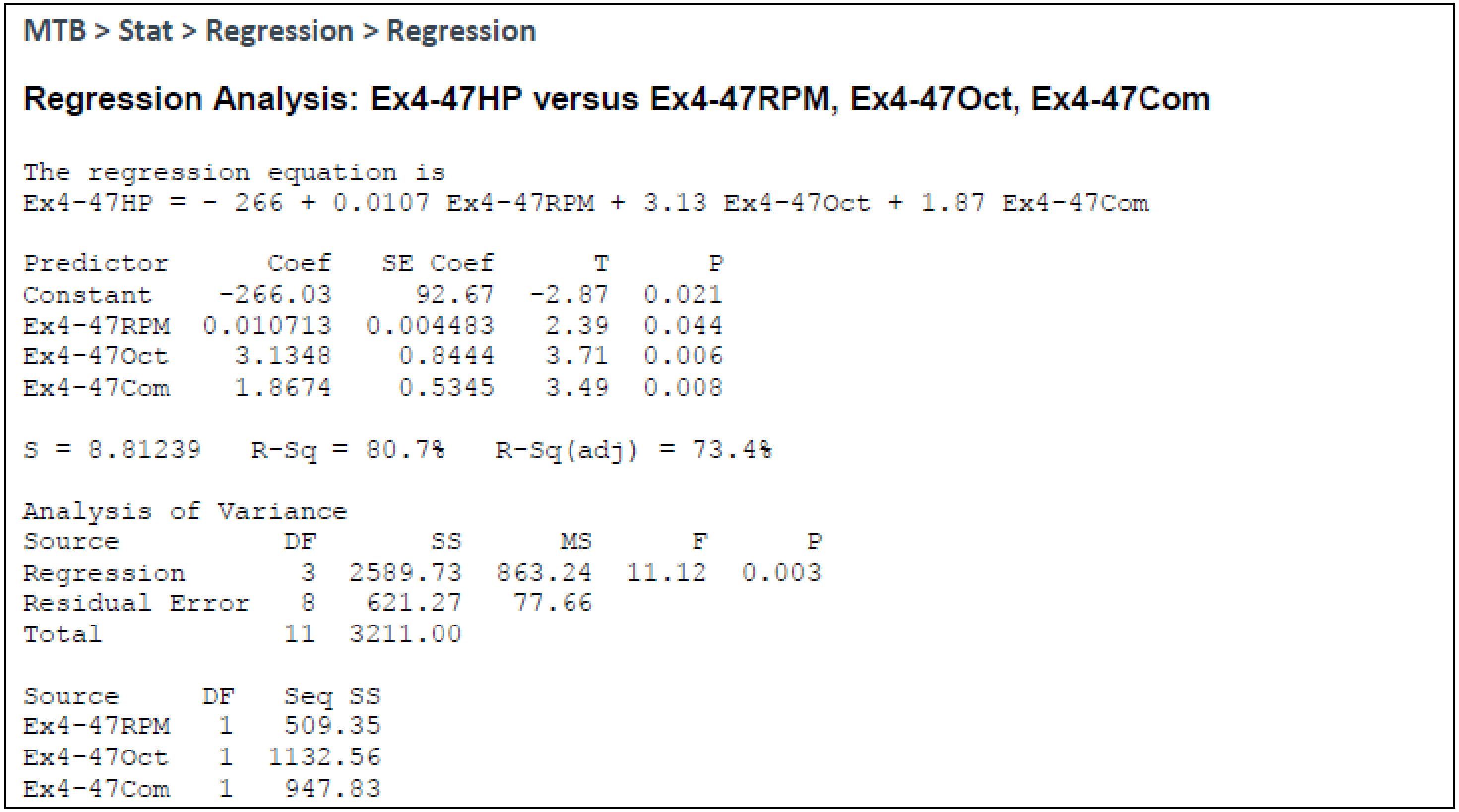
TABLE 4E.10

Automobile Engine Data for Exercise 4.47

Brake Horsepower	rpm	Road Octane Number	Compression
225	2,000	90	100
212	1,800	94	95
229	2,400	88	110
222	1,900	91	96
219	1,600	86	100
278	2,500	96	110
246	3,000	94	98
237	3,200	90	100
233	2,800	88	105
224	3,400	86	97
223	1,800	90	100
230	2,500	89	104

Regression Example (cont'd)

- Fit a regression model to the data



MTB > Stat > Regression > Regression

Regression Analysis: Ex4-47HP versus Ex4-47RPM, Ex4-47Oct, Ex4-47Com

```
The regression equation is
Ex4-47HP = - 266 + 0.0107 Ex4-47RPM + 3.13 Ex4-47Oct + 1.87 Ex4-47Com

Predictor        Coef   SE Coef      T      P
Constant      -266.03     92.67  -2.87  0.021
Ex4-47RPM    0.010713  0.004483   2.39  0.044
Ex4-47Oct      3.1348    0.8444   3.71  0.006
Ex4-47Com      1.8674    0.5345   3.49  0.008

S = 8.81239   R-Sq = 80.7%   R-Sq(adj) = 73.4%

Analysis of Variance
Source          DF       SS      MS      F      P
Regression       3  2589.73  863.24  11.12  0.003
Residual Error   8   621.27   77.66
Total           11  3211.00

Source      DF   Seq SS
Ex4-47RPM    1   509.35
Ex4-47Oct    1  1132.56
Ex4-47Com    1   947.83
```

Regression Example (cont'd)

- Test for overall significance
  - $F_0 = 11.12$
  - $F_{0.05,k,n-k-1} = F_{0.05,3,8} = 4.07$
- Test for individual term significance
  - *P*-values are all < 0.05
  - $R^2 = 0.807$
  - $R^2_{adj} = 0.734$

Wrap-Up

- Understand basics of uncertainty
- Understand random variables
- Use probability distributions to model uncertain random variables
- Understand basic statistical methods