v1.2

Probability and Statistics

Mathematical underpinnings of cost estimating

"God does not play dice with the cosmos." -Albert Einstein
"Do not presume to tell God what to do." -Niels Bohr

http://en.wikiquote.org/wiki/Quantum_mechanics

ICEAA CEBoK

v1.2

Acknowledgments

- ICEAA is indebted to TASC, Inc., for the development and maintenance of the Cost Estimating Body of Knowledge (CEBoK®)
  - ICEAA is also indebted to Technomics, Inc., for the independent review and maintenance of CEBoK®
- ICEAA is also indebted to the following individuals who have made significant contributions to the development, review, and maintenance of CostPROF and CEBoK ®
- Module 10 Probability and Statistics
  - Lead authors: Megan E. Dameron, Christopher J. Leonetti, Casey D. Trail
  - Assistant authors: Jessica R. Summerville, Jennifer K. Murrill, Sarah E. Grinnell
  - Senior reviewers: Richard L. Coleman, Kevin Cincotta, Fred K. Blackburn
  - Reviewers: Robyn Kane, Matthew J. Pitlyk, Maureen L. Tedford
  - Managing editor: Peter J. Braxton

ICEAA CEBoK

v1.2

Unit Index

Unit I - Cost Estimating

Unit II - Cost Analysis Techniques

Unit III - Analytical Methods

6. Basic Data Analysis Principles
7. Learning Curve Analysis
8. Regression Analysis
9. Cost and Schedule Risk Analysis
10. *Probability and Statistics*

Unit IV - Specialized Costing

Unit V - Management Applications

v1.2

Prob/Stat Overview

- Key Ideas
  - Probability and Statistics as "Flipsides"
  - Central Tendency and Dispersion
  - The Bell Curve
    - Normal (Gaussian) Distribution and the CLT
  - Inference

- Practical Applications
  - 6 Descriptive Statistics
    - Mean, Median, Mode, CV
  - 8 CER Development
    - t, F, R^2, CI, PI
  - 9 Modeling Uncertainty and Risk
    - Normal, Triangular, Lognormal

- Analytical Constructs
  - Counting and Fractions
    - Combinations and Permutations
    - Pascal's Triangle
  - Distributions and Calculus
    - pdfs and cdfs
    - Limits
    - (Maximum Likelihood) Estimators
  - Hypothesis Testing
    - Test statistic, critical values, significance

- Related Topics
  - Stochastic Processes
    - Markov Chains
    - Queueing Theory
  - Simulation
    - Discrete Event
    - Continuous
  - Data Analysis
  - Regression Analysis
  - Design of Experiments

Probability

Statistics

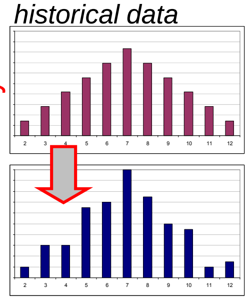
v1.2

Prob/Stat Within The Cost Estimating Framework

Past

Understanding your historical data

Probability

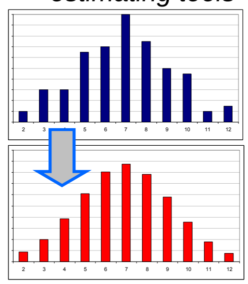


Observed outcomes are the result of the "cosmic roll of the dice"

Present

Developing estimating tools

Statistics

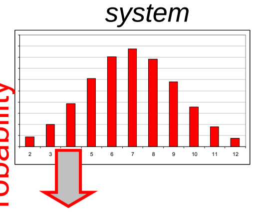


Statistics are used as "best guesses" for underlying population parameters

Future

Estimating the new system

Probability

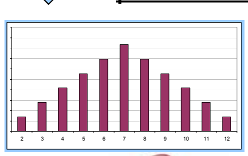


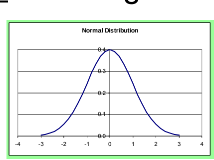
Fitted distributions are used to model subsequent outcomes

v1.2

Probability

- Probability is the mathematical study of the future; the chance of an event or outcome, or the range of possible outcomes
  - Divided into discrete and continuous probability
  - Encompasses a number of models for outcomes, called distributions, such as normal (or Gaussian), triangular, and many others
  - 11 There is a subset of probability called stochastic processes involving models that change over time





Note: This is a layman's definition

v1.2

Statistics

- Statistics is the mathematical study of the past; it involves describing outcomes, or inferring from outcomes what the underlying probability model might be
  - Divided into:
    - Descriptive statistics: involving the portrayal of data sets themselves and derived rates, averages, and the like
    - Inferential statistics: involving tests to determine if a given probabilistic model might apply, what the value of a parameter might be, or testing whether two sets of outcomes are differentiable
  - Divided into:
    - Parametric statistics: involving assumptions about the underlying model
    - Non-parametric statistics: involving few or no assumptions about the underlying model

Note: This is a layman's definition

v1.2

Role in Cost Estimating

6
- Descriptive statistics are used to describe and compare cost data

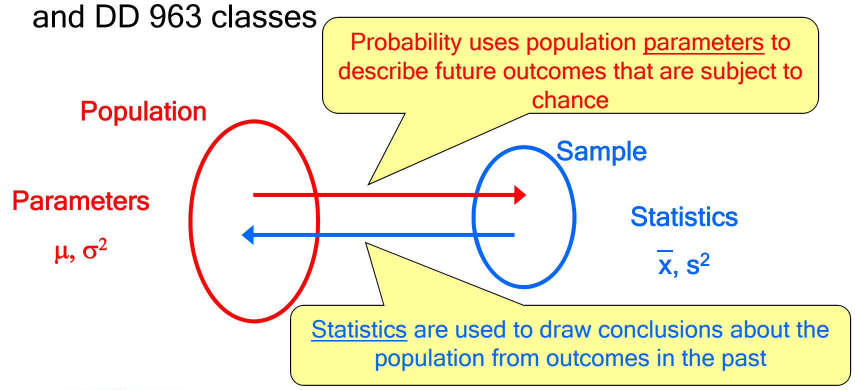
8
- Statistics provide the basis for the development of Cost Estimating Relationships (CERs) via regression
  - Inferential statistics are used to adjudge the goodness of those CERs

9
- Probability is used to quantify the uncertainty present in a cost estimate

v1.2

Definitions - Population / Sample

- A population consists of *all* members of a particular group, e.g., all (metaphysically possible) US Navy destroyers
- A sample is a subset of the population, e.g., DDG 51 and DD 963 classes



v1.2

Definitions - Random Variables

- A random variable takes on values that represent outcomes in the sample space
  - It cannot be fully controlled or exactly predicted
- Discrete vs. Continuous
  - A set is discrete if it consists of a finite or countably infinite number of values
    - e.g., number of circuits, number of test failures
    - e.g., {1,2,3, …} - the random variable can only have a positive integer (natural number) value
  - A set is continuous if it consists of an interval of real numbers (finite or infinite length)
    - e.g., time, height, weight
    - e.g., [-3,3] - the random variable can take on any value in this interval

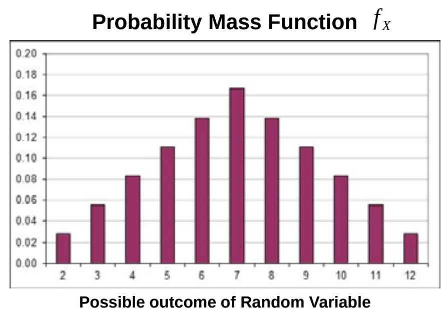
v1.2

Definitions - pmf

- Probability Mass Function (pmf)
  - The probability that the **discrete** random variable will take a value equal to *a* is the height of the bar at *a*.

$$P(X=a)=f_X(a)$$

  - The pmf accounts for all possible outcomes of the distribution
    - The sum of heights of all bars is 1.



The probability of a given outcome is the height of the corresponding histogram bar.

Tip: Discrete distributions are the exception, not the rule, in cost estimating

ICEAA CEBoK

Unit III - Module 10

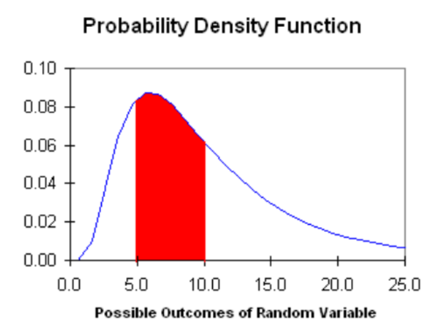
NEW!

v1.2

Definitions - pdf

- Probability Density Function (pdf)
  - The total area under the curve is 1
    - The probability that the random variable takes on some value in the range is 1 (100%)
  - The probability that the **continuous** random variable will take a value between *a* and *b* is the area under the curve between *a* and *b*

$$P(a \le X \le b)=\int_a^b p(x)dx$$



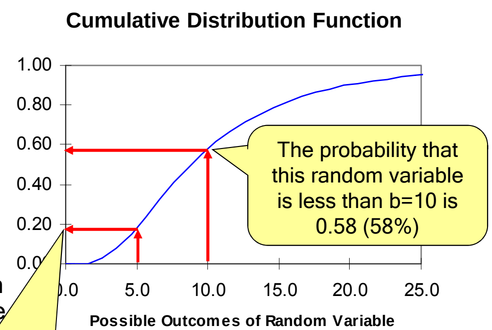
The probability that this random variable is between a=5 and b=10 is equal to the shaded area (40%)

ICEAA CEBoK

Unit III - Module 10

Definitions - cdf

- Cumulative Distribution Function (cdf)
  - This curve shows the probability that the random variable is *less* than x (a particular value of X)
  - The cdf reaches/ approaches 1
    - The probability that the random variable is less than the maximum value (may be infinite) is 1 (100%)



$$P(a \le X \le b) = F(b) - F(a)$$

$$= \int_{-\infty}^{b} p(x)dx - \int_{-\infty}^{a} p(x)dx = \int_{a}^{b} p(x)dx$$

Tip: The pdf shows the *shape* of the distribution; the cdf shows the *percentiles*

Distribution Properties

- Total probability = 1 (100%)

$$\sum_{i=1}^{\infty} p(x_i) = 1 \qquad \int_{-\infty}^{\infty} p(x)dx = 1$$

- cdf is sum of pmf / integral of pdf

$$F(x_j) = P(X \le x_j) = \sum_{i=1}^{j} p(x_i) \qquad F(x) = P(X \le x) = \int_{-\infty}^{x} p(t)dt$$

- pmf is delta / pdf is derivative of cdf

$$p(x_j) = P(X = x_j) = \sum_{i=1}^{j} p(x_i) - \sum_{i=1}^{j-1} p(x_i) = F(x_j) - F(x_{j-1})$$

$$p(x) = \lim_{t \to 0} \frac{F(x+t) - F(x)}{t} = F'(x)$$

v1.2

Measures of Central Tendency

Where is the "center" of the distribution?

v1.2

Mean

6

- The expected value of a distribution (population mean), is calculated as the sum (integral) of a random variable's possible values multiplied by the probability that it takes on those values

1

$$E(X)=\sum x_i p(x_i) \qquad E(X)=\int xp(x)dx$$

$$E(X)=2\left(\frac{1}{36}\right)+3\left(\frac{2}{36}\right)+\ldots+7\left(\frac{6}{36}\right)+\ldots+11\left(\frac{2}{36}\right)+12\left(\frac{1}{36}\right)=7$$

$$E(X)=\int_{-\infty}^{\infty}\frac{x}{\sigma\sqrt{2\pi}}e^{-\frac{(x-\mu)^2}{2\sigma^2}}dx=\mu\int_{-\infty}^{\infty}\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{(x-\mu)^2}{2\sigma^2}}dx+\int_{-\infty}^{\infty}\frac{(x-\mu)}{\sigma\sqrt{2\pi}}e^{-\frac{(x-\mu)^2}{2\sigma^2}}dx$$

$$=\mu+\frac{\sigma}{\sqrt{2\pi}}e^{-\frac{(x-\mu)^2}{2\sigma^2}}\Bigg|_{-\infty}^{\infty}=\mu$$

v1.2

Note that the median may not be in the distribution

Median

6

- The <u>median</u> of a distribution is the value that exactly divides the distribution (pdf) into equal halves (middle value or average of two middle values); "robust" to extreme values

2

$\sum_{x_i \le m} p(x_i) \ge \frac{1}{2}$ $\sum_{x_i \ge m} p(x_i) \ge \frac{1}{2}$ $\int_{-\infty}^{m} p(x)dx = \frac{1}{2}$

$$\int_{-\infty}^{\mu} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx + \int_{\mu}^{\infty} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx = 1$$

$\int_{m}^{\infty} p(x)dx = \frac{1}{2}$

$$\int_{-\infty}^{\mu} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx = \int_{\mu}^{\infty} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx = \frac{1}{2}$$

$m = \mu$

The median of a Normal distribution - as with any symmetric distribution with finite mean - is its mean

v1.2

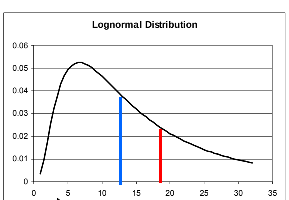
Mean, Median, and Skew

6

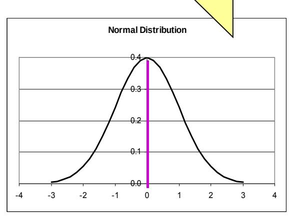
- The mean and the median are equal if the distribution is <u>symmetric</u>
- Unequal means and medians are an indication of <u>skewness</u>

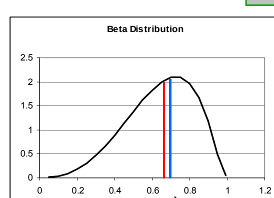
4

20



Median = Mean
Symmetric





Median < Mean
Skew(ed) Right

Median > Mean
Skew(ed) Left

v1.2

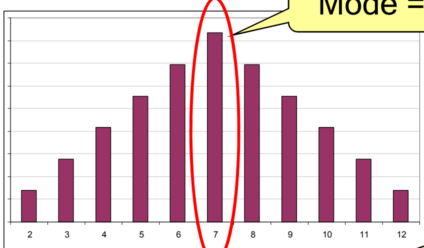
Mode

6

- The mode of a distribution is the most frequent value in the distribution (discrete) or the value with the greatest probability density (continuous)

3

Mode = 7



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

$$p'(x) = -\frac{(x-\mu)}{\sigma^3\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

pdf is increasing to the left of the mean and decreasing to the right of the mean

The mode of a Normal distribution - as with any *unimodal* symmetric distribution - is its mean

v1.2

Measures of Dispersion

How "spread out" is the distribution?

v1.2

Variance / Standard Deviation

6

- The variance of a distribution is the measure of the "spread" of the distribution about its mean – the second "moment"

std dev of about 2.42

$$Var(X) = E\left((X-\mu)^2\right) = \sum (x_i - \mu)^2 p(x_i)$$

$$Var(X) = 25\left(\frac{1}{36}\right) + 16\left(\frac{2}{36}\right) + \ldots + 0\left(\frac{6}{36}\right) + \ldots + 16\left(\frac{2}{36}\right) + 25\left(\frac{1}{36}\right) = \frac{35}{6} = 5.8\overline{3}$$

$$Var(X) = E\left((X-\mu)^2\right) = \int (x-\mu)^2 p(x)dx$$

$$Var(X) = \int_{-\infty}^{\infty} \frac{(x-\mu)^2}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx = -\frac{\sigma(x-\mu)}{\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}\Bigg|_{-\infty}^{\infty} + \sigma^2 \int_{-\infty}^{\infty} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx = \sigma^2$$

The variance of a distribution is the square of σ, its standard deviation

ICEAA CEBoK

v1.2

Coefficient of Variation

6

- The Coefficient of Variation (CV) expresses the standard deviation as a percent of the mean

$$CV = \frac{\sigma}{\mu}$$

Tip: Low CV indicates less dispersion, i.e., a tighter distribution

6

8

5

Large CVs indicate that the mean is a poor representation of the distribution
  - Specify distribution using complete set of parameters, if possible
  - Include other parameters, such as variance
- CV is *invariant to scaling*.
  - e.g., CV{1,2,3}=CV{100,200,300}
- Not to be confused with the CV regression statistic

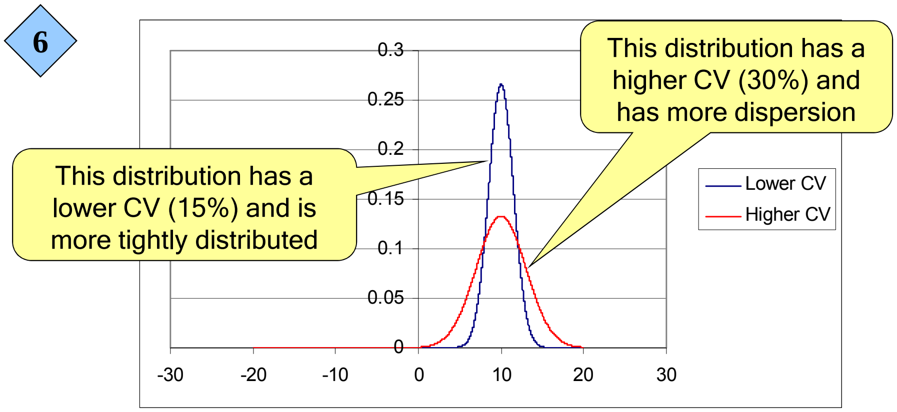
ICEAA CEBoK

Dispersion and CV

16

- These two distributions have the same mean, but different standard deviations

6



Probability Distributions

- Normal
- Student's t
- Lognormal
- F
- Triangular
- Bernoulli
- Relationships between Distributions

Normal (Gaussian)

v1.2

- The normal distribution, or "bell-shaped curve," is the most prevalent distribution
  - Many naturally-occurring phenomena have a normal distribution, such as the height of people
- The normal distribution is used in many statistical tests and applications
- The normal distribution is symmetric about the mean

Unit III - Module 10

30

Normal - Parameters

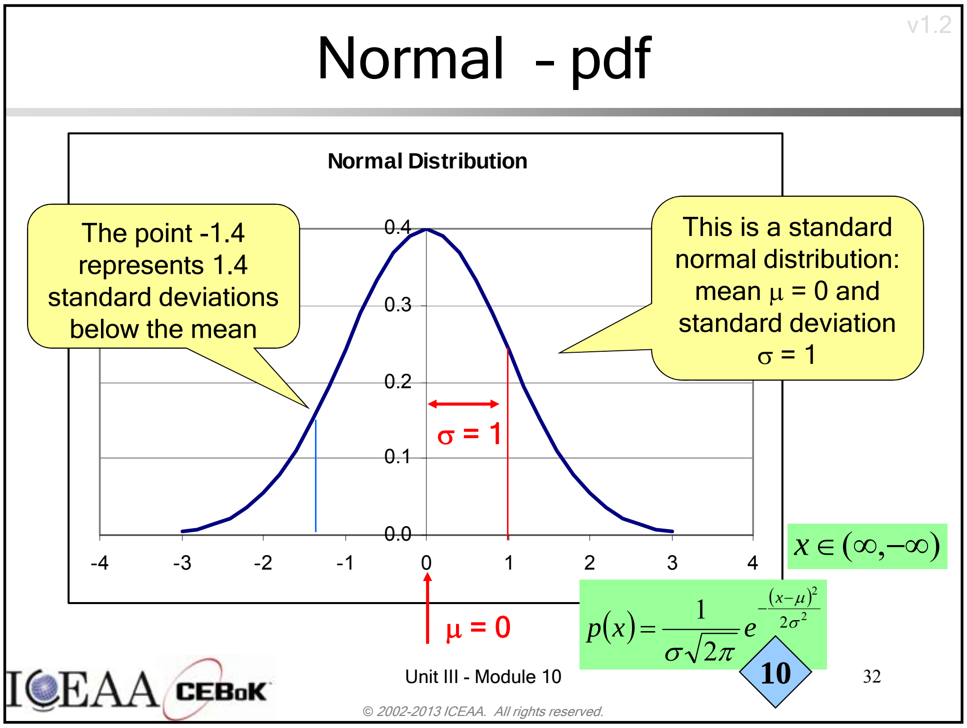
v1.2

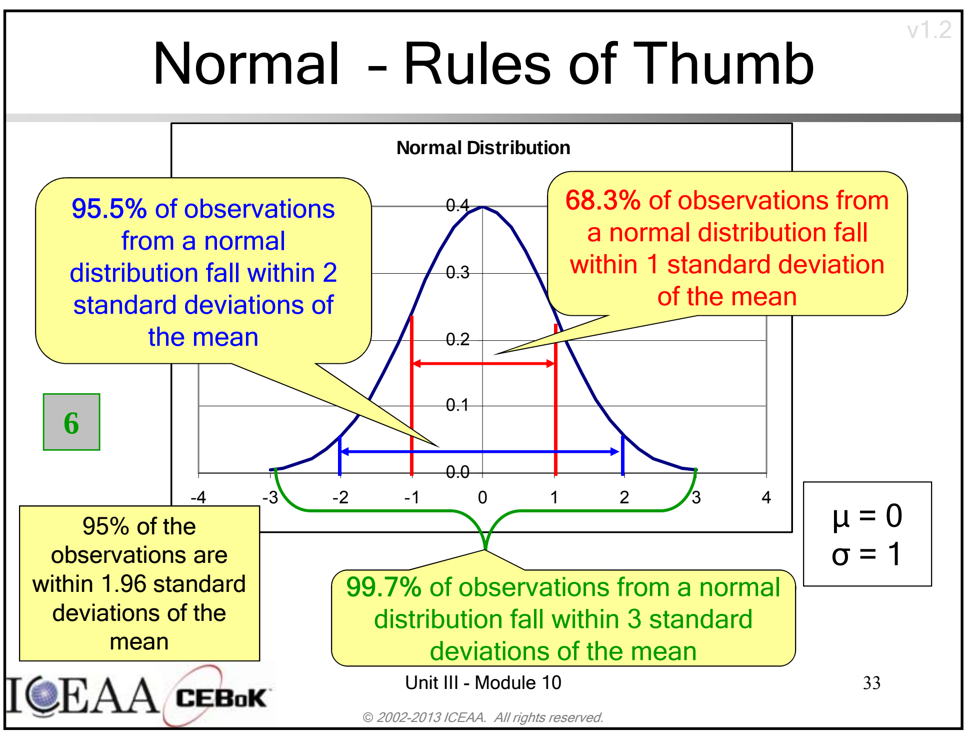
- The normal distribution has two parameters:
  - The mean of the distribution, μ
  - The standard deviation, σ

- If X is a random variable with a normal distribution, we write $X \sim N(\mu, \sigma^2)$
- A standard normal is a normal distribution with $\mu = 0$ and $\sigma = 1$ and is denoted $Z \sim N(0,1)$
  - If $X \sim N(\mu, \sigma^2)$, then $\frac{(X-\mu)}{\sigma} \sim N(0,1)$

Unit III - Module 10

31


v1.2
Normal - pdf
Normal Distribution
The point -1.4 represents 1.4 standard deviations below the mean
This is a standard normal distribution: mean μ = 0 and standard deviation σ = 1
σ = 1
μ = 0
0.4
0.3
0.2
0.1
0.0
-4 -3 -2 -1 0 1 2 3 4
$x \in (\infty, -\infty)$
$p(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$
10
ICEAA CEBoK
Unit III - Module 10
32
© 2002-2013 ICEAA. All rights reserved.


v1.2
Normal - Rules of Thumb
Normal Distribution
95.5% of observations from a normal distribution fall within 2 standard deviations of the mean
68.3% of observations from a normal distribution fall within 1 standard deviation of the mean
6
0.4
0.3
0.2
0.1
0.0
-4 -3 -2 -1 0 1 2 3 4
μ = 0
σ = 1
95% of the observations are within 1.96 standard deviations of the mean
99.7% of observations from a normal distribution fall within 3 standard deviations of the mean
ICEAA CEBoK
Unit III - Module 10
33
© 2002-2013 ICEAA. All rights reserved.

v1.2

Central Limit Theorem (CLT)

7

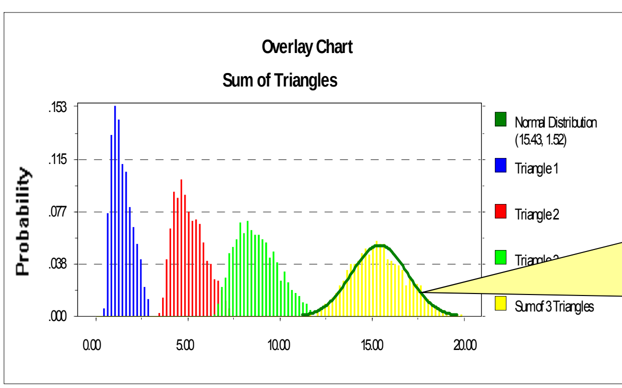
- The sum of a large number of *independent*, *identically distributed* (*iid*) random variables from a population with finite mean and standard deviation approaches a normal distribution
  - Sample size required depends on the parent distribution, but as a rule of thumb, distributions approach normal by n = 30
- Correlation: As long as the sum is not dominated by a few large, highly correlated elements, the CLT will still hold

Normality of Work Breakdown Structures, M. Dameron, J. Summerville, R. Coleman, N.St. Louis, Joint ISPA/SCEA Conference, June 2001.

v1.2

CLT - Example

- The graph below shows 3 triangular distributions and the sum of the 3 triangles



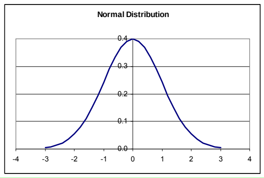
The sum of three independent triangular distributions tests as being normally distributed

"Normality of Work Breakdown Structures," M. Dameron, J. Summerville, R. Coleman, N. St. Louis, Joint ISPA/SCEA Conference, June 2001.

v1.2

Normal Distribution Overview

Distribution



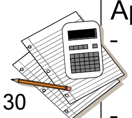
$x \in (\infty, -\infty)$

$$p(x)=\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{(x-\mu)^2}{2\sigma^2}} \qquad F(x)=\Phi(x)=\int_{-\infty}^{x}\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{(t-\mu)^2}{2\sigma^2}}dt$$

Parameters and Statistics

- Mean = μ
- Variance = σ^2
- Skewness = 0
- 2.5th percentile = μ - 1.96σ
- 97.5th percentile = μ + 1.96σ

Key Facts



- If X ~ N(μ, σ^2), then $\frac{(X-\mu)}{\sigma}$ ~ N(0,1) ("standard" normal)
- Central Limit Theorem holds for n ≥ 30
- 68.3/95.5/99.7 Rule
- Limiting case of t distribution
- Exponential of normal is lognormal
- Dist: NORM.DIST(x, mean, stddev, cum)
  - cum = TRUE for cdf and FALSE for pdf
- Inv cdf: NORM.INV(prob, mean, stddev)

Applications

10

- Central Limit Theorem
  - Approximation of distributions
  - Limiting case of t-distribution
- Regression Analysis
  - Assumed error term
- Distribution of cost
  - Default distribution
- Distribution of risk
  - Symmetric risks and uncertainties

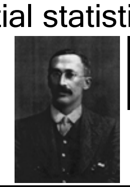
ICEAA CEBoK

v1.2

Student's t Distribution

- The t distribution is similar to the standard normal, but is flatter with more area in the tails
  - Parameter is the <u>degrees of freedom</u>, n
    - In t tests, n is directly related to sample size
    - Note that in regression, the degrees of freedom will be (n-1)-k, where n is the number of data points and k is the number of independent variables
  - Symmetric about the mean
- As the degrees of freedom increase, the t distribution approaches normal
- The t distribution is very important for confidence intervals and hypothesis testing (inferential statistics), which will be explained in later slides

8

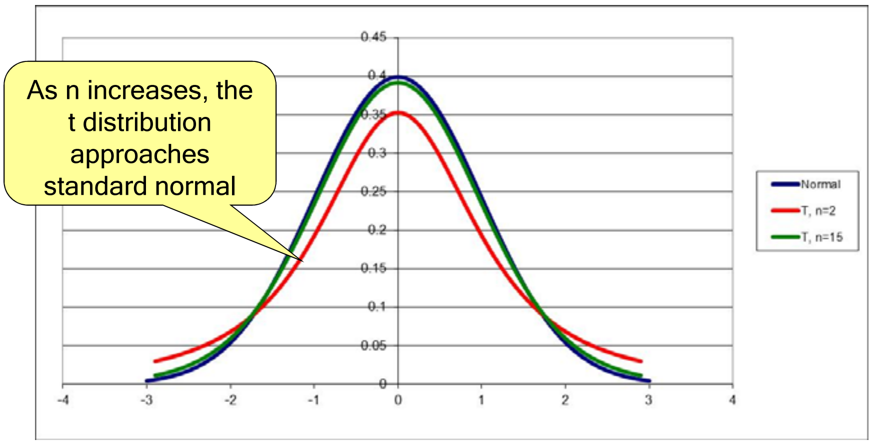


GUINNESS

ICEAA CEBoK

t Distribution - pdf

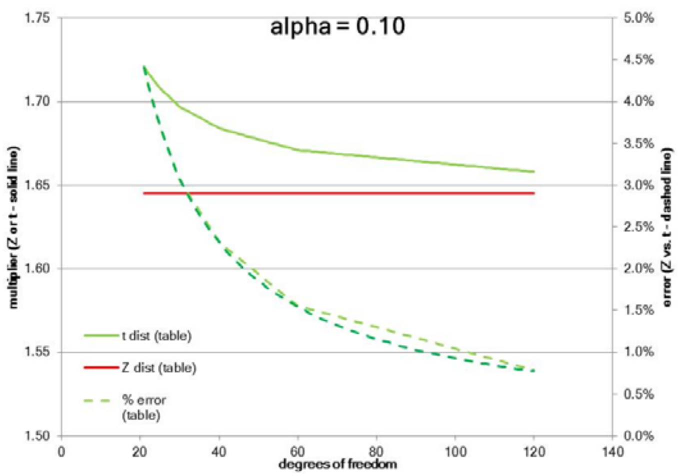
v1.2



Warning: Never approximate a t distribution with a standard normal!

Danger of Approximating t Distribution with a Normal Distribution

v1.2

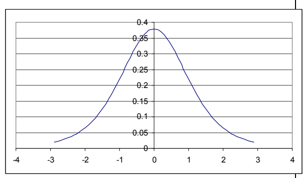


NEW!

t Distribution Overview

v1.2

Distribution



$x \in (-\infty, \infty)$

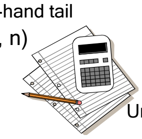
$$p(x) = \frac{\Gamma[(n+1)/2]}{\sqrt{n\pi}\,\Gamma(n/2)[1+(x^2/n)]^{(n+1)/2}}$$

Parameters and Statistics

- Degrees of freedom = n
- Mean = 0
- Variance = $\frac{n}{n-2}$
- Skewness = 0

Key Facts

- As n approaches infinity, the t distribution approaches Standard Normal
- The t is distributed as $t = \frac{N(0,1)}{\chi(n)/\sqrt{n}}$
- Excel
  - cdf = T.DIST(x, n, tails)
    - Default is left-hand tail, use .2T for two tails, .RT for right-hand tail
  - Inv cdf = T.INV(prob, n)



Applications

- Confidence Intervals
  - Mean of Normal variates
- Regression Analysis
  - Significance of individual coefficients
- Hypothesis Testing

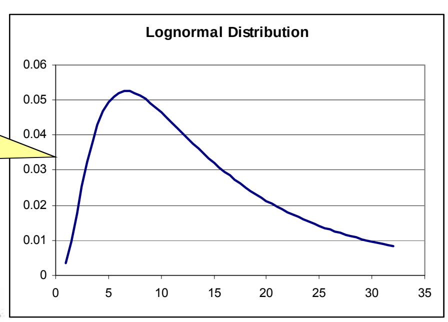
ICEAA CEBoK

Lognormal Distribution

v1.2

- The lognormal distribution :
  - Formed by raising e to the power of ("exponentiating") a normal random variable. $Y \sim N(\mu, \sigma^2) \Rightarrow e^Y$ is lognormal.
  - *Not* the log of a normal. Rather, a variable is lognormal if its (natural) log is normal. Also, the (natural) log of a lognormal is normal.
  - The mean of the *related normal* distribution, μ
  - The standard deviation of the *related normal* distribution, σ

The lognormal distribution is skewed right



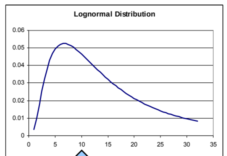
Tip: If the distribution of X is lognormal, then the natural log (ln) of X is normally distributed.

ICEAA CEBoK

v1.2

Lognormal Distribution Overview

- Distribution



$x \in [0, \infty)$

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

10

Parameters and Statistics

- Median = e^{μ}
- Std Deviation of ln X = σ
- Mean = $e^{\mu + \sigma^2/2}$
- Variance = $e^{2\mu+\sigma^2}\left(e^{\sigma^2} - 1\right)$

Key Facts

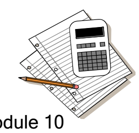
- If X has a lognormal distribution, then ln(X) has a normal distribution
- For small standard deviations, the normal approximates the lognormal distribution
  - For CVs < 25%, this holds

10

- Excel
  - Cdf = LOGNORM.DIST(x, mean, stddev)
  - Inv cdf = LOGNORM.INV(prob, mean, stddev)

Applications

- Risk Analysis



v1.2

Triangular Distribution

- The triangular distribution has three parameters: minimum, mode, and maximum 18
- Can be symmetric or skewed
- Often used in risk analysis
  - Especially useful for eliciting and quantifying expert opinions
  - Almost never found anywhere else

9

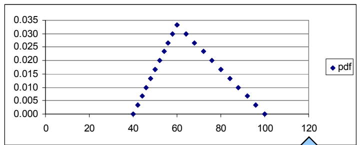
"Do Not Sum 'Most Likely' Costs," Stephen A. Book, IDA/OSD CAIG Cost Symposium, May 1992.

Triangular Distribution Overview

v1.2

19

Distribution



$$p(x)=\begin{cases} 2(x-a)/[(b-a)(c-a)] & a \le x \le c \\ 2(b-x)/[(b-a)(b-c)] & c \le x \le b \end{cases}$$

10

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

Parameters and Statistics

- Min = a
- Max = b
- Mode = c ($a \le c \le b$)
- Mean = $\frac{a+b+c}{3}$
- Variance = $\frac{a^2+b^2+c^2-ab-ac-bc}{18}$

Key Facts

- Excel
  - pdf and cdf calculations can be handled as they are listed in the above formulas
- A symmetrical triangle approximates a normal when

$$a=\mu-\sqrt{6}\sigma, c=\mu, b=\mu+\sqrt{6}\sigma$$

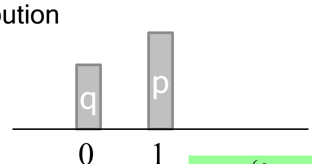
Applications

- Risk Analysis
  - SME Input

Bernoulli Distribution Overview

v1.2

Distribution



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

$$F(x)=\begin{cases} 0 & x<0 \\ q & 0 \le x < 1 \\ 1 & x \ge 1 \end{cases}$$

Parameters and Statistics

- Min = 0
- Max = 1
- Mean = p
- Variance = pq

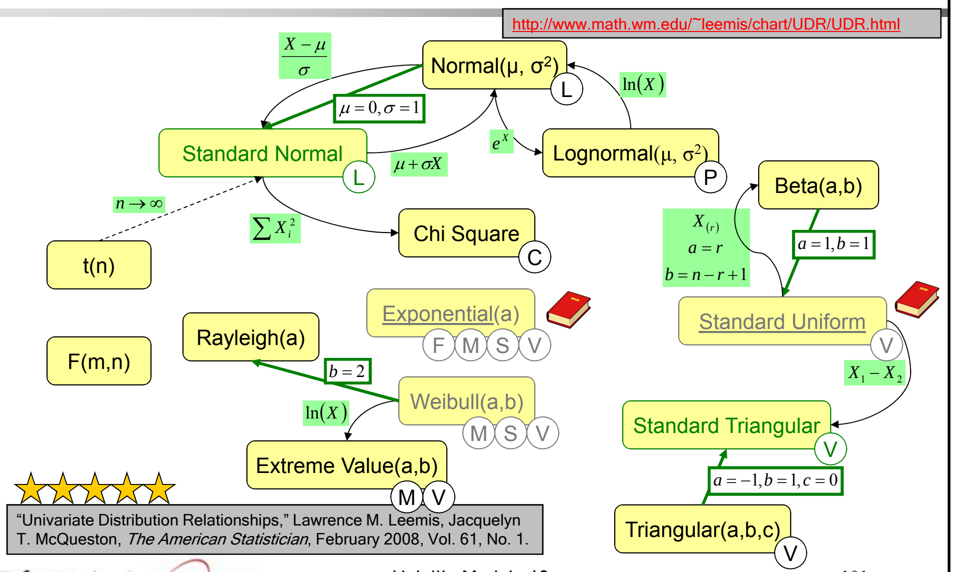
Key Facts

- Excel
  - pdf and cdf calculations can be handled as they are listed in the above formulas
- The sum of n Bernoullis is Binomial (n,p)

Applications

- Risk Analysis
  - Discrete Risks
  - X = Cf * Bernoulli
  - p = Pf

Relationships Between Distributions



"Univariate Distribution Relationships," Lawrence M. Leemis, Jacquelyn T. McQueston, *The American Statistician*, February 2008, Vol. 61, No. 1.

Hypothesis Testing

- One Tail vs. Two Tail
- Statistical Significance
- Test Statistics
- Critical Values
- P-Values
- Confidence Intervals

v1.2

Hypothesis Test

- Hypothesis tests are often used to test for differences between population groups
- Two hypotheses are proposed:
  - H_0 is the null hypothesis
    - H_0 is presumed to be true unless the data is proven to contradict the null hypothesis
  - H_1 is the alternative hypothesis
    - H_1 may only be accepted with statistical evidence contradicting the null hypothesis

"innocent until proven guilty"

"beyond a reasonable doubt"

v1.2

Examples of Hypotheses

- Test to see if two populations have different means

  H_0: $\mu_1 = \mu_2$ (the means are the same)
  H_1: $\mu_1 \neq \mu_2$ (the means are different)
- Test to see if two populations have different standard deviations

  H_0: $\sigma_1 = \sigma_2$ (the std devs are the same)
  H_1: $\sigma_1 \neq \sigma_2$ (the std devs are different)
- Test to see if two populations are identically distributed

  H_0: $f(x) = g(x)$ (the distributions are the same)
  H_1: $f(x) \neq g(x)$ (the distributions are different)

t test

F test

Chi Square or K-S test

v1.2

One Tail vs. Two Tail

- A hypothesis can be one-tailed or two-tailed
- A <u>one-tailed test</u> makes an assumption about the *direction* of difference

10

  - E.g., H_0: $\mu_1 = \mu_2$
    H_1: $\mu_1 > \mu_2$
- A <u>two-tailed test</u> makes no assumption about the direction of difference
  - E.g., H_0: $\mu_1 = \mu_2$
    H_1: $\mu_1 \neq \mu_2$

We will look at this case in our main example.

ICEAA CEBoK

v1.2

Statistical Significance

- First, we must choose a <u>level of significance</u>, denoted α
- α is the probability of incorrectly rejecting the null hypothesis
  - This is called a <u>Type I error</u>
  - Typical significant levels are α = 0.05 and α = 0.10
- The customary level of significance is α = 0.05
  - Rejecting H_0 at an α = 0.05 level of significance means that there is less than a 5% probability that H_0 is true

Convicting an innocent person

Tip: α = 0.05 is the most common level of significance. α = 0.10 is occasionally used. Any other values are very rare

ICEAA CEBoK

Test Statistic

- A test statistic is a function of the sample data
  - Calculated under the assumption that the null hypothesis is true

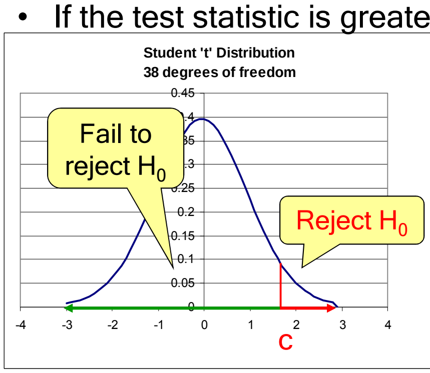
11

- The decision to accept or reject the null hypothesis is based on the value of the test statistic
- Different types of hypothesis tests will have different test statistics (many of these will be discussed on later slides)

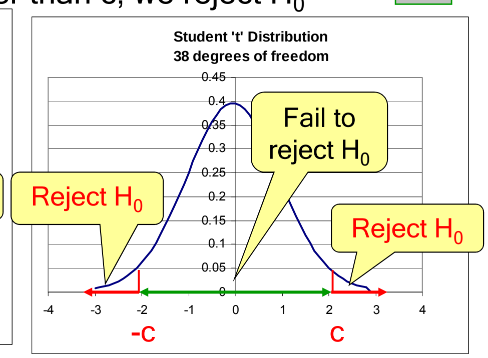
Critical Value

- A critical value c is such that the probability of getting a test statistic greater than c (in absolute value) is equal to α for a one tailed test and $\alpha/2$ for a two-tailed test
- If the test statistic is greater than c, we reject H_0

12



$P(t>c) = \alpha$



$P(t<-c) = \alpha/2$ $P(t>c) = \alpha/2$

Example Problem

- Suppose we have historical cost growth factors from a set of DoD programs and a set of NAVAIR programs
- We wish to see if cost growth for NAVAIR programs differs from DoD-wide growth
- The hypotheses are:

  H_0: $\mu_N = \mu_D$ (the means are the same)

  H_1: $\mu_N \neq \mu_D$ (the means are different)
- This is a two-tailed test as we are making no assumptions as to whether or not NAVAIR has higher or lower growth than DoD

"NAVAIR Cost Growth Study," R.L. Coleman, M.E. Dameron, C.L. Pullen, J.R. Summerville, D.M. Snead, 34th DoDCAS and ISPA/SCEA 2001.

Example Problem Data

- Suppose we have the following cost growth factors (CGF) for DoD and NAVAIR programs
- Average DoD CGF = 1.19
- Average NAVAIR CGF = 1.33

But is this a "real" difference, or did it happen by chance?

DoD	NAVAIR
1.26	1.26
1.44	1.92
0.96	1.64
0.93	1.83
1.26	1.85
0.88	1.03
1.10	1.08
1.23	1.44
0.76	1.60
1.75	1.24
1.80	1.04
1.56	1.21
1.24	1.11
1.51	1.31
0.93	1.47
0.49	1.25
1.29	1.11
0.76	1.15
1.21	1.11
1.50	0.90

Example Test Statistic

v1.2

- For each different type of hypothesis test, there is a corresponding test statistic
- In our example problem, we will be using the two-sample t-test for means
  - Let $X_n \sim N(\mu_x, \sigma^2)$ and $Y_m \sim N(\mu_y, \sigma^2)$ where X and Y are independent
  - Let S_x^2 and S_y^2 be the two sample variances
  - Let S_p^2 be the pooled variance, where

$$S_p^2 = \frac{(n-1)S_X^2 + (m-1)S_Y^2}{n+m-2}$$

  - Then, $T = \dfrac{\overline{X} - \overline{Y} - (\mu_X - \mu_Y)}{S_p\sqrt{\dfrac{1}{n} + \dfrac{1}{m}}}$

has a t distribution with (n + m - 2) degrees of freedom

Example Test Statistic Calculation

v1.2

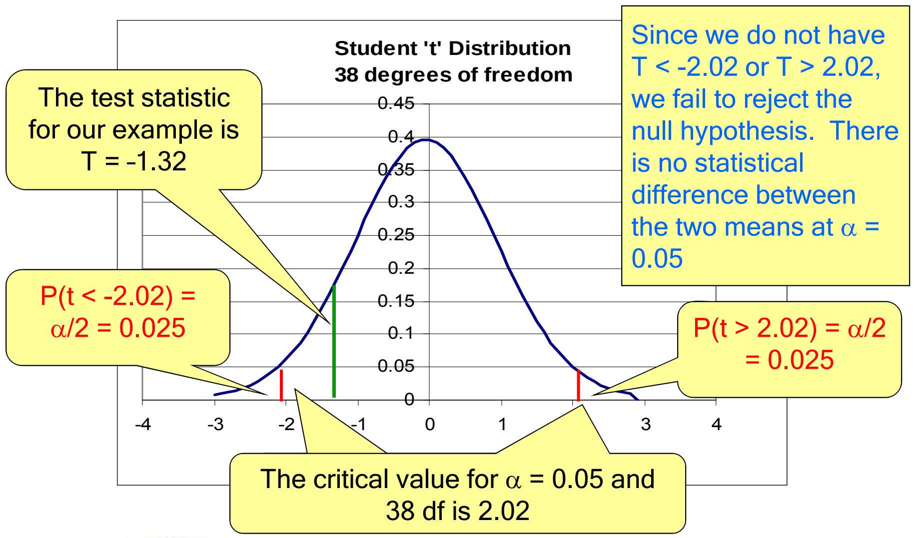
- Using our example problem data, we get the following:
- $\overline{X} = 1.19$, $\overline{Y} = 1.33$
- $S_x^2 = 0.12$, $S_y^2 = 0.09$
- $S_p^2 = \dfrac{(20-1)(0.12) + (20-1)(0.09)}{20 + 20 - 2}$

  $= 0.11$
- If H_0 is true, then $(\mu_x - \mu_y) = 0$
- So, under H_0,

$$T = \frac{(1.19 - 1.33) - 0}{\sqrt{0.11(\frac{1}{20} + \frac{1}{20})}} = -1.32$$

DoD	NAVAIR
1.26	1.26
1.44	1.92
0.96	1.64
0.93	1.83
1.26	1.85
0.88	1.03
1.10	1.08
1.23	1.44
0.76	1.60
1.75	1.24
1.80	1.04
1.56	1.21
1.24	1.11
1.51	1.31
0.93	1.47
0.49	1.25
1.29	1.11
0.76	1.15
1.21	1.11
1.50	0.90

Example Critical Values



Example t Test Using Excel

- In Excel, use the Data Analysis Add-In to run a t test

	DoD	*NAVAIR*
Mean	1.193094	1.327704
Variance	0.117941	0.089239
Observations	20	20
Pooled Variance	0.10359	
Hypothesized Mean Difference	0	
df	38	
t Stat	-1.322571	
P(T<=t) one-tail	0.096942	
t Critical one-tail	1.685953	
P(T<=t) two-tail	0.193883	
t Critical two-tail	2.024394	

Tip: The Excel default significance level is α = 0.05

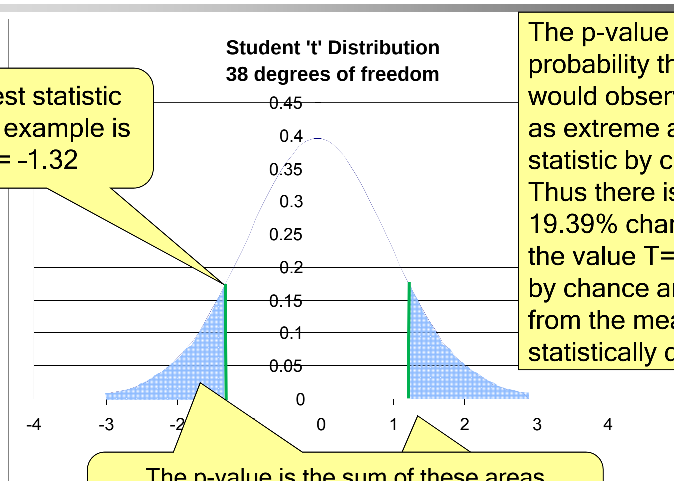
test statistic for our example

Critical value for a one-tailed test. In this case, we would still fail to reject H_0

Critical value for our two-tailed test. Fail to reject the null hypothesis.

Warning: Results of macros do not update if your data change!

Example p-Values



The test statistic for our example is T = -1.32

The p-value is the probability that we would observe a value as extreme as the test statistic by chance. Thus there is a 19.39% chance that the value T=-1.32 was by chance and not from the means being statistically different.

The p-value is the sum of these areas indicating P(t<-1.32)+P(t>1.32)= 2(0.096942)=0.193883

Example p-Values Using Excel

- The p-value is the smallest level of significance for which H_0 could have been rejected

	DoD	*NAVAIR*
Mean	1.193094	1.327704
Variance	0.117941	0.089239
Observations	20	20
Pooled Variance	0.10359	
Hypothesized Mean Difference	0	
df	38	
t Stat	-1.322571	
P(T<=t) one-tail	0.096942	
t Critical one-tail	1.685953	
P(T<=t) two-tail	0.193883	
t Critical two-tail	2.024394	

For a two-tailed test, we could reject the null hypothesis of equal means at α = 0.194. But then there would still be a 19.4% probability that we incorrectly rejected H_0.

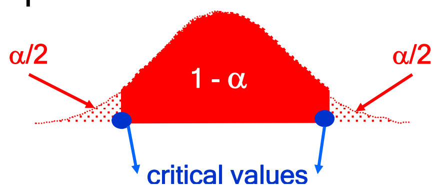
Warning: Choose the level of significance ahead of time and stick with it!

v1.2

Confidence Intervals

13

- A confidence interval (CI) suggests to us that we are $(1-\alpha)\cdot 100\%$ confident that the true parameter value is contained within the calculated range*
  - The range is calculated using the estimated parameter value
- Confidence intervals can be calculated for a variety of different parameters and distributions



* Note this statement provides a general sense of what a confidence interval does for us in concise language for ease of understanding. The specific statistical interpretation is that if many independent samples are taken where the levels of the predictor variable are the same as in the data set, and a $(1-\alpha)*100\%$ confidence interval is constructed for each sample, then $(1-\alpha)\cdot 100\%$ of the intervals will contain the true value of the parameter.

v1.2

Example CI Formula

- Let us suppose that the Cost Growth Factors in our example problem are normally distributed
- We will find a 95% confidence interval for the *average* DoD cost growth
- The formula is

$$\left(\bar{y} - t_{\alpha/2,n-1}\frac{s}{\sqrt{n}}, \mu, \bar{y} + t_{\alpha/2,n-1}\frac{s}{\sqrt{n}}\right)$$

Example CI Calculation

v1.2

- Going back to the data from our example problem, we find:
  - $\bar{y}$ = mean DoD Cost Growth Factor = 1.19
  - s = standard deviation of DoD data = 0.34
  - n = sample size = 20
- We want a 95% CI, so we have $\alpha = 0.05$
- Now, we need $t_{\alpha/2,\, n-1}$
  - We can find this on a table or by using the TINV function in Excel
  - $t_{0.025,\, 19} = 2.093$

Tip: Use α with T.INV.2T or 1- α/2 with T.INV.

Example CI Result

v1.2

- So, our confidence interval is

9

$$\left(1.19 - 2.093\frac{0.34}{\sqrt{20}}, \mu, 1.19 + 2.093\frac{0.34}{\sqrt{20}}\right)$$

$$(1.03, \mu, 1.35)$$

Roughly speaking, we are 95% certain that the true value of the DoD Cost Growth Factor mean is between 1.03 and 1.35

v1.2

Prob/Stat Summary

- A solid understanding of probability and statistics is vital to both cost and risk analysis
- Descriptive statistics are used to characterize, describe, and compare the data
  - Central Tendency – mean, median, mode
  - Dispersion – variance, standard deviation, coefficient of variation
- Inferential statistics are used to draw inferences from the data
  - Testing multiple population groups for differences in means, variances, distributions, etc.
  - Confidence intervals around estimates