

Review of Probability and Statistics

OR-651
Spring 2008

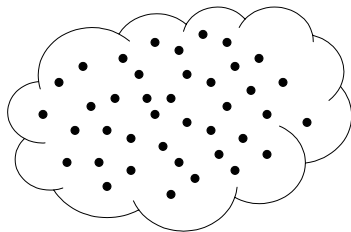
Review of Probability and Statistics

- Outline:
 - Statistics overview
 - Probability overview
 - Confidence intervals

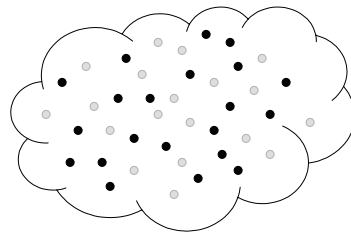
Statistics Overview

Population vs. Sample

- **Population** (universe) is the totality of all things under consideration.
 - E.g., all members of the US Navy
- A **Sample** is a portion of the population selected for analysis
 - E.g., those sailors on a certain ship whose SSNs end in 7.



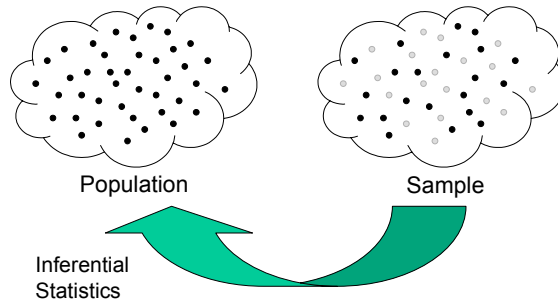
Population



Sample

Descriptive vs. Inferential Statistics

- **Descriptive Statistics** are those methods involving the **collection, presentation and characterization** of a set of data in order to properly describe the features of that data.
- **Inferential Statistics** are those methods that facilitate the **estimation** of population characteristics based on **sample** results.

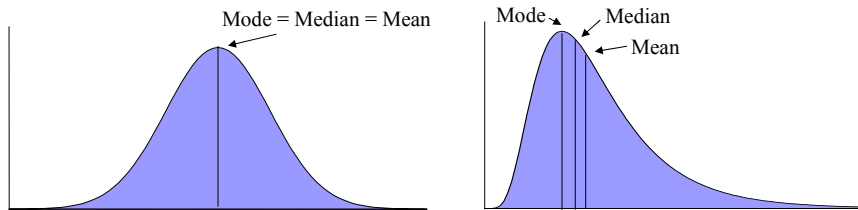


Parameter vs. Statistic

- A **Parameter** is a summary measure that describes a characteristic of a **population**.
 - E.g., ~52% of all humans are female
- A **Statistic** is a summary measure that describes a characteristic from a **sample**.
 - E.g., 5% of sailors sampled have used drugs in the last four weeks
- The objective of **Statistics** is to make **inferences** (predictions, decisions) about a **population** based upon information contained in a **sample**.
 - Textbook definition
- The objective of **Statistics** is to make **estimates** about the cost of a **weapon system** based upon information contained in **analogous systems**.
 - DoD Cost Analyst's definition

Measures of Central Tendency

- These statistics describe the “middle region” of the sample.
 - Mean
 - The arithmetic average of the data set.
 - Median
 - The “middle” of the data set.
 - Mode
 - The value in the data set that occurs most frequently.
- These are almost never the same, unless you have a perfectly symmetric, unimodal population.



Mean

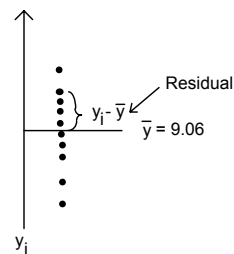
- The **Sample Mean** (\bar{y}) is the arithmetic average of a data set.
- It is used to estimate the population mean, (μ).
- Calculated by taking the sum of the observed values (y_i) divided by the number of observations (n).

Historical Transmogripher Average Unit Production Costs

System	FY06\$K
1	22.2
2	17.3
3	11.8
4	9.6
5	8.8
6	7.6
7	6.8
8	3.2
9	1.7
10	1.6

$$\bar{y} = \frac{\sum_{i=1}^n y_i}{n} = \frac{y_1 + y_2 + \dots + y_n}{n}$$

$$\bar{y} = \frac{22.2 + 17.3 + \dots + 1.6}{10} = \$9.06K$$



Median

- The **Median** is the **middle** observation of an ordered (from low to high) data set
- Examples:
 - 1, 2, 4, 5, 5, 6, 8
 - Here, the middle observation is 5, so the median is 5
 - 1, 3, 4, 4, 5, 7, 8, 8
 - Here, there is no “middle” observation so we take the average of the two observations at the center

$$\text{Median} = \frac{4 + 5}{2} = 4.5$$

- Unlike the Mean, the Median is resistant to extreme outliers
 - 1, 2, 4, 5, 5, 6, 8, 1000 (same as first example, but with one additional extreme observation)
 - But note that the Median is STILL just 5!

Mode

- The **Mode** is the value of the data set that occurs **most frequently**
- Example:
 - 1, 2, 4, 5, 5, 6, 8
 - Here the Mode is 5, since 5 occurred twice and no other value occurred more than once
- Data sets can have more than one mode, while the mean and median have one unique value
 - 1, 2, 2, 2, 5, 7, 7, 7, 8, 10
 - This data set has two modes...2 and 7
- Data sets can also have NO mode, for example:
 - 1, 3, 5, 6, 7, 8, 9
 - Here, no value occurs more frequently than any other, therefore no mode exists

Dispersion Statistics

- The **Mean**, **Median** and **Mode** by themselves are not sufficient descriptors of a data set
- Example:
 - Data Set 1: 48, 49, 50, 51, 52
 - Data Set 2: 5, 15, 50, 80, 100
- Note that the Mean and Median for both data sets are identical, but the data sets are glaringly different!
- The difference is in the **dispersion** of the data points
- Dispersion Statistics we will discuss are:
 - Range
 - Variance
 - Standard Deviation

Range

- The **Range** is simply the difference between the **smallest** and **largest** observation in a data set
- Example
 - Data Set 1: 48, 49, 50, 51, 52
 - Data Set 2: 5, 15, 50, 80, 100
- The Range of data set 1 is $52 - 48 = 4$
- The Range of data set 2 is $100 - 5 = 95$
- So, while both data sets have the same mean and median, the dispersion of the data, as depicted by the range, is much smaller in Data Set 1

Variance

- The **Sample Variance**, s^2 , measures the amount of variability of the sample data relative to their mean
- As shown below, the variance is the “average” of the squared deviations of the observations about their mean

$$s^2 = \frac{\sum (y_i - \bar{y})^2}{n - 1}$$

- The *sample variance* is used to **estimate** the actual *population variance*, σ^2

$$\sigma^2 = \frac{\sum (y_i - \mu)^2}{N}$$

Standard Deviation

- The Variance is not a “common sense” statistic because it describes the data in terms of squared units
- The **Sample Standard Deviation**, s , is simply the square root of the **sample variance**

$$s = \sqrt{\frac{\sum (y_i - \bar{y})^2}{n - 1}}$$

- The **sample standard deviation** is used to **estimate** the actual *population standard deviation*, σ

$$\sigma = \sqrt{\frac{\sum (y_i - \mu)^2}{N}}$$

Standard Deviation

- The **sample standard deviation**, s , is measured in the same units as the data from which it is being calculated

System	FY06\$K	$y_i - \bar{y}$	$(y_i - \bar{y})^2$
1	22.2	13.1	172.7
2	17.3	8.2	67.9
3	11.8	2.7	7.5
4	9.6	0.5	0.3
5	8.8	-0.3	0.1
6	7.6	-1.5	2.1
7	6.8	-2.3	5.1
8	3.2	-5.9	34.3
9	1.7	-7.4	54.2
10	1.6	-7.5	55.7
Average	9.06		

$$s^2 = \frac{\sum (y_i - \bar{y})^2}{n-1}$$

$$= \frac{172.7 + 67.9 + \dots + 55.7}{10-1}$$

$$= \frac{399.8}{9} = 44.4 (\$K^2)$$

$$s = \sqrt{s^2} = \sqrt{44.4 (\$K^2)}$$

$$= 6.67 (\$K)$$

- This number, \$6.67K, represents the “average” distance of each data point from the sample mean

Coefficient of Variation

- For a given data set, the standard deviation is \$100,000.
- Is that good or bad? It depends...
 - A standard deviation of \$100K for a task estimated at \$5M would be very good indeed.
 - A standard deviation of \$100K for a task estimated at \$100K is clearly useless.
- What constitutes a “good” standard deviation?
- The “goodness” of the standard deviation is not its value per se, but rather what percentage the standard deviation is of the estimated value.
- The **Coefficient of Variation (CV)** is defined as the “average” percent distance of each data point from the sample mean.
- The CV is the ratio of the standard deviation to the mean.

$$CV = \frac{s_y}{y}$$

Coefficient of Variation

- In the first example, the CV is $\$100\text{K}/\$5\text{M} = 2\%$
- In the second example, the CV is $\$100\text{K}/\$100\text{K} = 100\%$
- These values are unitless and can be readily compared.
- *The CV is the “average” percent estimating error for the population when using \bar{y} as the estimator.*
- *Or, the CV is the “average” percent estimating error when estimating the cost of future tasks.*
- Calculate the CV from our previous transmogrifier cost database:
 - $CV = \$6.67\text{K}/\$9.06\text{K} = 73.6\%$
- *Therefore, for subsequent observations we would expect to be off on “average” by 73.6% when using \$9.06K as the estimated cost.*

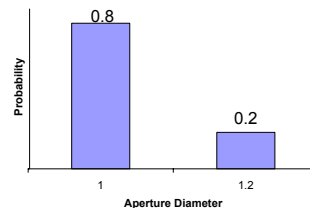
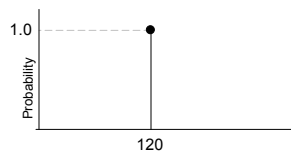
Probability Overview

Probability Distributions

- There are a large variety of probability distributions that are typically used in cost analysis applications
- Some of the more commonly used distributions include the following:
 - Deterministic (no distribution)
 - Discrete (few choices)
 - Uniform (lowest, highest)
 - Triangular (lowest, most likely, highest)
 - Normal (μ, σ)
 - Lognormal (μ, σ)

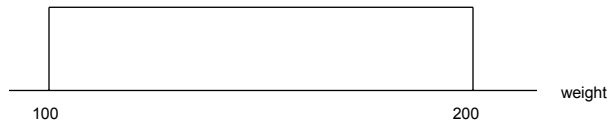
Probability Distributions

- Deterministic
 - One choice is to have no distribution at all
 - Example:
 - Weight = 120 lbs
 - If a deterministic value is used, then it is assumed that no uncertainty exists
- Discrete
 - A discrete distribution is one in which only certain outcomes, with associated probabilities, are allowed
 - Example:
 - Weight = 120 lbs with probability 0.8, or
 - Weight = 200 lbs with probability 0.2



The Uniform Distribution

- One might choose to model a random variable with a uniform distribution if all that is known is the minimum possible and maximum possible values of the random variable, with all values in between being equally likely
- This distribution is most often used to model the input values of cost models
 - For example, structure weight may be as low as 100 lbs or as high as 200 lbs, with all possibilities in between equally likely



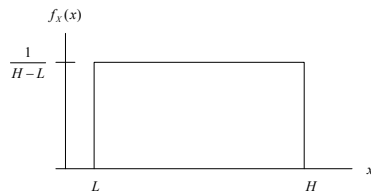
The Uniform Distribution

- The PDF of a uniform distribution is:

$$f_X(x) = \frac{1}{H-L} \quad \text{if } L \leq x \leq H$$

where $-\infty < L < H < \infty$.

- The uniform PDF and its mean and variance are illustrated below:

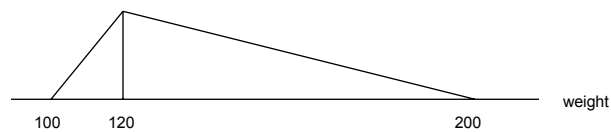


$$E(X) = \frac{(L+H)}{2}$$

$$Var(X) = \frac{1}{12}(H-L)^2$$

The Triangular Distribution

- One might choose to model a random variable with a triangular distribution if all that is known is the lowest possible (L), most likely (M), and highest possible (H) values of the random variable
- This distribution is most often used to model the input values of cost models
 - For example, structure weight is most likely to be about 120 lbs, but may be as low as 100 lbs or as high as 200 lbs



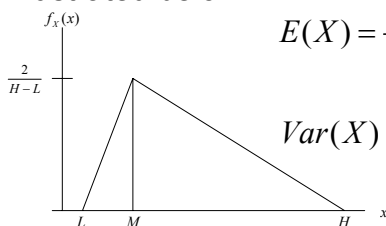
The Triangular Distribution

- The PDF of a triangular distribution is:

$$f_x(x) = \begin{cases} \frac{2(x-L)}{(H-L)(M-L)} & \text{if } L \leq x < M \\ \frac{2(H-x)}{(H-L)(H-M)} & \text{if } M \leq x < H \end{cases}$$

where $-\infty < L < M < H < \infty$.

- The triangular PDF and its mean and variance are illustrated below:

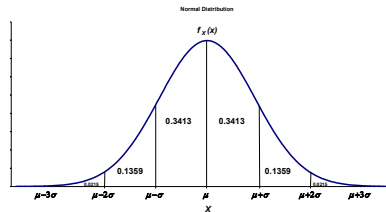


$$E(X) = \frac{(L + M + H)}{3}$$

$$Var(X) = \frac{1}{18} \left((M-L)(M-H) + (H-L)^2 \right)$$

The Normal Distribution

- One might model a random variable with a normal distribution having mean μ and standard deviation σ if one expected the distribution to be symmetric, bell-shaped, and if it is expected that almost all observations would fall within $\pm 3\sigma$ of the mean



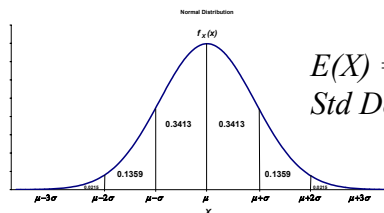
The Normal Distribution

- The normal distribution is defined by the following PDF:

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

where $-\infty < x < \infty$, $\sigma > 0$ and μ is unrestricted

- Also known as the Gaussian distribution, the normal PDF is uniquely defined by the parameters μ and σ



$$E(X) = \mu$$

$$Std Dev(X) = \sigma$$

The Normal Distribution

- As with any probability distribution, the area under the curve, $f_X(x)$, is defined as 1.0:

$$P(-\infty < X < \infty) = \int_{-\infty}^{\infty} f_X(x) dx = 1.0$$

- The normal distribution is symmetric about its mean. It also has well-defined probabilities associated with various distances away from the mean, for example:

$$P(\mu - \sigma \leq X \leq \mu + \sigma) = \int_{\mu - \sigma}^{\mu + \sigma} f_X(x) dx = 0.6826$$

$$P(\mu - 2\sigma \leq X \leq \mu + 2\sigma) = \int_{\mu - 2\sigma}^{\mu + 2\sigma} f_X(x) dx = 0.9544$$

$$P(\mu - 3\sigma \leq X \leq \mu + 3\sigma) = \int_{\mu - 3\sigma}^{\mu + 3\sigma} f_X(x) dx = 0.9973$$

The Lognormal Distribution

- The lognormal distribution is closely related to the normal distribution
 - If X is a non-negative random variable, and $Y = \ln(X)$ follows a normal distribution, then X is said to have a lognormal distribution

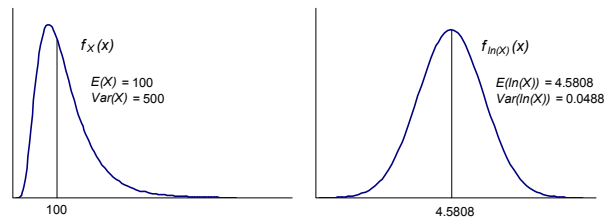
The Lognormal Distribution

- The PDF of a lognormally distributed random variable X is:

$$f_X(x) = \frac{1}{\sqrt{2\pi}\sigma_Y x} e^{-\frac{1}{2}\left[\frac{(\ln(x)-\mu_Y)^2}{\sigma_Y^2}\right]}$$

where $0 < x < \infty$, $\sigma_Y > 0$, $\mu_Y = E(\ln(X))$ and $\sigma_Y^2 = \text{Var}(\ln(X))$

- The lognormal PDF and its related normal PDF are illustrated below:



The Lognormal Distribution

- If the mean and variance of the related normal distribution are known, then the mean and variance of the lognormal distribution can be calculated as follows:

$$E(X) = \mu_X = e^{\mu_Y + \frac{1}{2}\sigma_Y^2}$$

$$\text{Var}(X) = \sigma_X^2 = e^{2\mu_Y + \sigma_Y^2} (e^{\sigma_Y^2} - 1)$$

The Lognormal Distribution

- However, when using the lognormal distribution to model cost, we typically do not have values of μ_Y and σ_Y^2 , but they can be calculated from $E(X) = \mu_X$ and $Var(X) = \sigma_X^2$ as follows:

$$\mu_Y = E(\ln X) = \frac{1}{2} \ln \left[\frac{(\mu_X)^4}{(\mu_X)^2 + \sigma_X^2} \right]$$

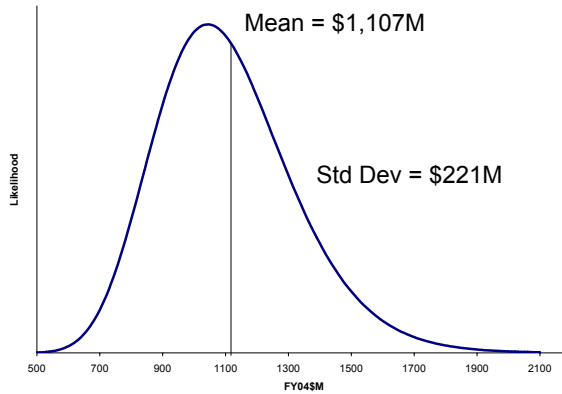
$$\sigma_Y^2 = Var(\ln X) = \ln \left[\frac{(\mu_X)^2 + \sigma_X^2}{(\mu_X)^2} \right]$$

Example Uses of Distributions

Probability Distribution	Example
Normal	Cost factor
Lognormal	Non-linear cost model
Deterministic	Aperture diameter
Discrete	Launch vehicle
Uniform	Labor rates, man-hours
Triangular	Software lines of code

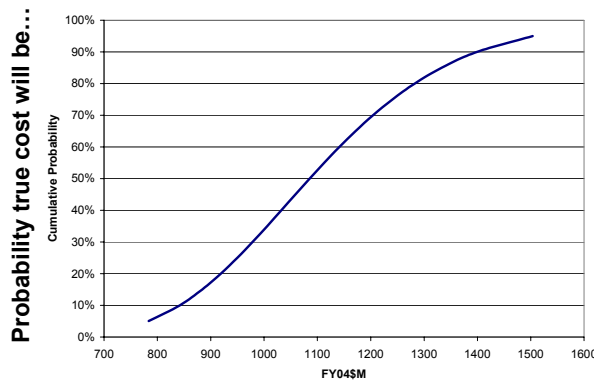
Probability Density Function

- Describes the shape and moments of the cost distribution
- The mean is the weighted average cost
- The standard deviation measures the spread of the distribution



Cumulative Distribution Function

- Describes the quantiles (percentiles) of the cost distribution
- Can also be represented in a table of percentiles

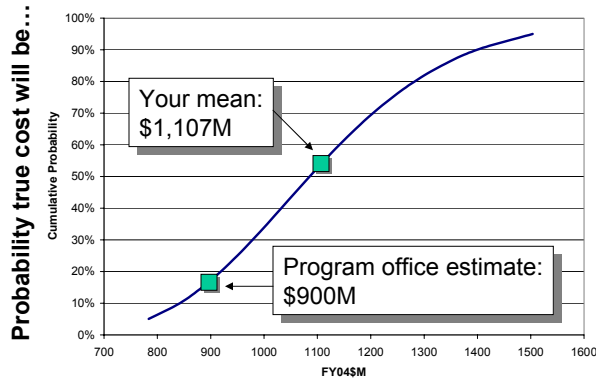


Percentiles	
5%	\$ 784
10%	\$ 842
15%	\$ 884
20%	\$ 919
25%	\$ 950
30%	\$ 978
35%	\$ 1,006
40%	\$ 1,032
45%	\$ 1,059
50%	\$ 1,086
55%	\$ 1,113
60%	\$ 1,141
65%	\$ 1,172
70%	\$ 1,204
75%	\$ 1,241
80%	\$ 1,282
85%	\$ 1,333
90%	\$ 1,399
95%	\$ 1,503

...less than or equal to this number

Cumulative Distribution Function

- Since the probability distribution represents *your* cost estimating uncertainty, you can compare anyone else's estimate to yours
- Those that fall at the lower percentiles are unlikely to be high enough!



Suppose a program office gives you an estimate of \$900M.

According to what you know about the system, there is only about an 18% chance that \$900M will be enough!

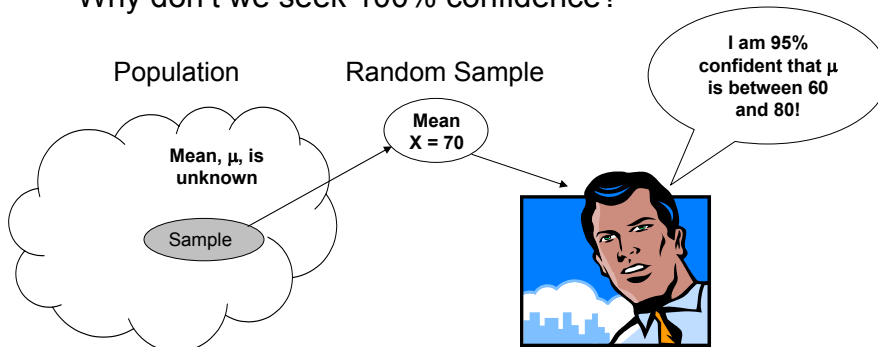
Confidence Intervals

Introduction

- Estimating confidence intervals is one of the most effective forms of statistical inference.
- In polling, we hear things like:
 - “Based on a sample of 600, 45% of Americans think the President is doing a good job...these results have a margin of error of ± 3 percentage points.”
- What this really means is that, statistically, one can conclude, with a certain degree of confidence (usually 90% or 95%), that the true population approval rating is $45\% \pm 3\%$ (or 42% to 48%) based on this sample of 600 Americans.

Estimation Process

- We use confidence intervals to estimate the bounds of the true population mean based on a sample.
- We don't really know the true population mean, but we are, say, 95% sure that we have it bounded.
- Why don't we seek 100% confidence?

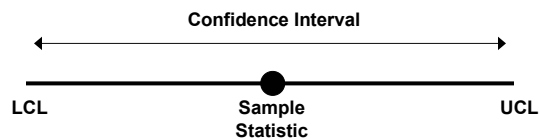


Confidence Interval Estimation

- Provides a range of values within which we think the true parameter lies, with a specified degree of confidence, based on information contained in a sample.
- But, since our estimate of the true population parameter is based on a sample, we can never be 100% sure (unless we sample the entire population).

Confidence Interval Estimation

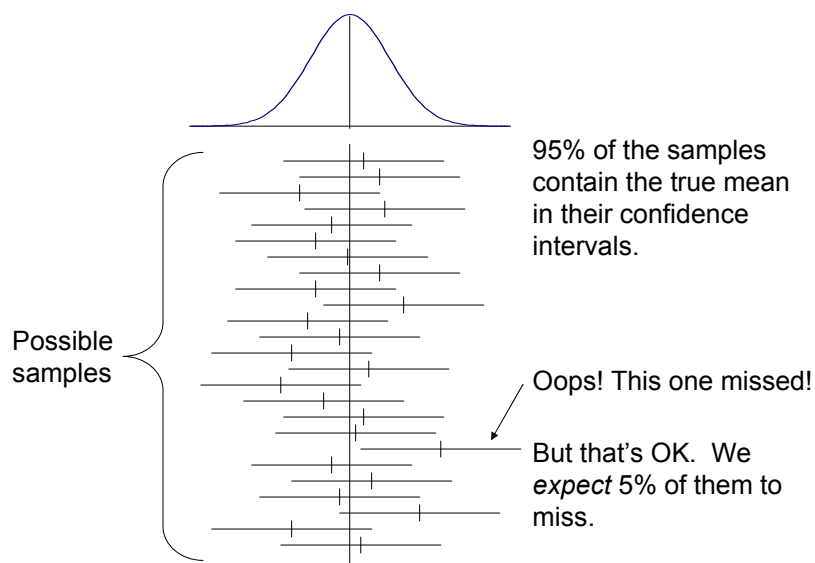
- We start a confidence interval estimate by specifying a probability that the true population parameter will fall somewhere within that interval.
 - E.g., 90%, 95%
- Then, given a sample statistic, we determine the necessary width of that interval, centered on the sample statistic, and bounded by a lower confidence limit and an upper confidence limit



Interpretation

- A 95% confidence interval estimate is interpreted as follows:
 - If all possible samples of size n are taken, and their sample means are computed, then 95% of them include the true population mean somewhere within the interval around their sample means and only 5% of them do not.
 - Because only one sample is selected in practice, and the true mean is unknown, we never know for sure whether the specific interval we've calculated includes the population mean.
 - However, we can state that we have 95% confidence that we have selected a sample whose confidence interval does include the population mean.

Interpretation



Confidence Limits for the Mean

- In general, a population mean, μ , is equal to the sample average \pm some error.

$$\mu = \bar{X} \pm \text{Error}$$

- We measure the error as:

$$\text{Error} = \pm(\bar{X} - \mu)$$

- If the population has a normal distribution with known σ , then:

$$Z = \frac{\bar{X} - \mu}{\sigma_{\bar{X}}} = \frac{\text{Error}}{\sigma_{\bar{X}}}$$

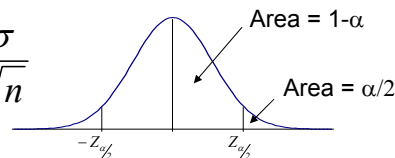
$$\text{Error} = Z\sigma_{\bar{X}} = Z \frac{\sigma}{\sqrt{n}}$$

$$\mu = \bar{X} \pm Z \frac{\sigma}{\sqrt{n}}$$

Calculating Confidence Limits

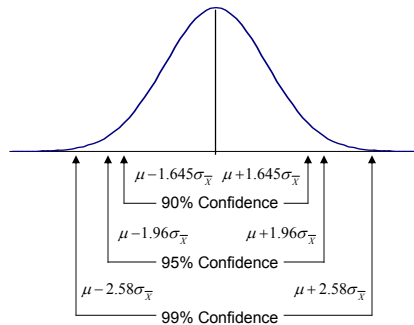
- The confidence interval is a function of the desired probability, the sample size, and the variance of the population distribution.
- The $(1-\alpha)$ confidence interval for a mean with a known σ is:

$$\bar{X} - Z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + Z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$$



- Note: α is the probability that the parameter is **not** within the interval.

Confidence Intervals



- This graphic shows a 90% CI, a 95% CI, and a 99% CI.

Example

- Suppose we desire a 90% CI for a sample of size $n=1000$, with $\bar{X} = 20$ and $\sigma = 5$ (known in advance).

$$(1 - \alpha)\% \text{ CI} = \bar{X} \pm Z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$$

$$1 - \alpha = 90\% \rightarrow \alpha = 0.1 \rightarrow \alpha/2 = 0.05$$

$$\bar{X} = 20 \rightarrow \sigma = 5 \rightarrow n = 1000$$

$$Z_{\alpha/2} = Z_{0.05} = 1.645 \text{ (from standard normal tables)}$$

$$90\% \text{ CI} = 20 \pm 1.645 \frac{5}{\sqrt{1000}} = 20 \pm 0.26 = (19.74, 20.26)$$

- Interpretation: We have 90% confidence that the true mean is somewhere between 19.74 and 20.26.

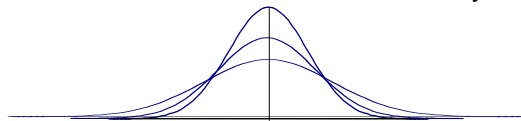
Confidence Intervals: σ unknown

- In practice, it is unusual that we would know the true value of σ .
- So...the previous analysis was used as a stepping stone to get us to this point...estimating a confidence interval when s is unknown, using only the sample statistics \bar{X} and s .
- In this case, we replace the *normal* distribution with the *Student's t* distribution.

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

The *Student's t* Distribution

- Recall that if $\bar{X} \sim \text{Normal}(\mu, \sigma/\sqrt{n})$, then $Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$ has a standard normal distribution.
- But, if σ is unknown, we estimate it with s , meaning the overall uncertainty is larger than if σ were known.
- At the same time, the larger the sample size, n , the less uncertainty we have about μ .
- So, the *t* distribution is really a *family* of distributions that have many of the same properties as the standard normal distribution, except that it has fatter tails for smaller values of n .
- And, as n gets large, the *t* distribution is equivalent to the standard normal distribution.
- When $n \geq 120$, the two distributions are virtually identical.



Degrees of Freedom

- $t_{\alpha/2, n-1}$ gives a critical value for a distribution whose mean is zero, and is based on $n-1$ degrees of freedom.
- What do we mean by “degrees of freedom?”
 - Recall that the sample variance is calculated as $\frac{\sum (x_i - \bar{X})^2}{n-1}$
 - Thus, in order to compute s^2 , we first need to know \bar{X} .
 - Therefore, we can say that only $n-1$ of the sample values are free to vary (because since we know \bar{X} , the n^{th} sample must be fixed). Therefore, there are $n-1$ degrees of freedom.
 - Example: If $\bar{X} = 2$, $X_1 = 1$, and $X_2 = 2$, then X_3 must be equal to 3 (it cannot vary).

$$\bar{X} = \frac{1+2+X_3}{3} = 2 \Leftrightarrow X_3 = (2)(3) - 1 - 2 = 3$$

Example

- Suppose we desire a 95% CI for a sample of size $n=25$, with $\bar{X} = 50$ and $s = 8$.

$$(1-\alpha)\% \text{ CI} = \bar{X} \pm t_{\alpha/2, n-1} \frac{s}{\sqrt{n}}$$
$$1-\alpha = 95\% \rightarrow \alpha = 0.05 \rightarrow \alpha/2 = 0.025$$
$$\bar{X} = 50 \rightarrow s = 8 \rightarrow n = 25$$
$$t_{\alpha/2, n-1} = t_{0.025, 24} = 2.0639 \text{ (from standard t tables)}$$

$$95\% \text{ CI} = 50 \pm 2.0639 \frac{8}{\sqrt{25}} = 50 \pm 3.30 = (46.69, 53.30)$$

- Interpretation: We have 95% confidence that the true mean is somewhere between 46.69 and 53.30.

Summary

- Statistics overview
- Probability overview
- Confidence intervals