Sampling Error

Data Science for Quality Management: Sampling Distributions, Error and Estimation

with **Wendy Martin**

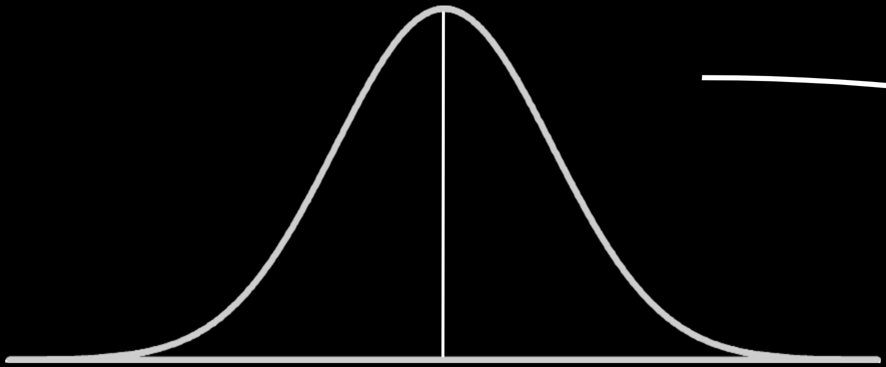
Learning objectives:

Describe the concept of sampling error

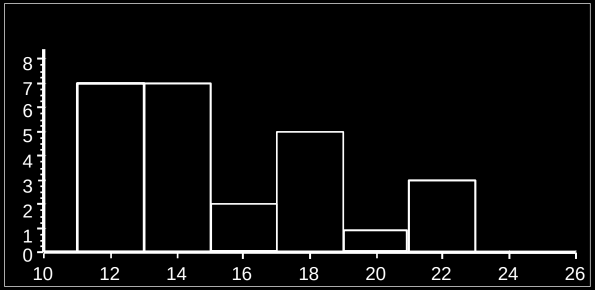
Create a vector of normally distributed random numbers

Sampling Distributions and Estimation

Are you ***really*** interested in the sample?



Population



Sample

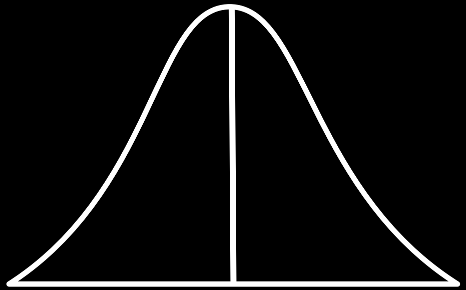
Parameters and Statistics

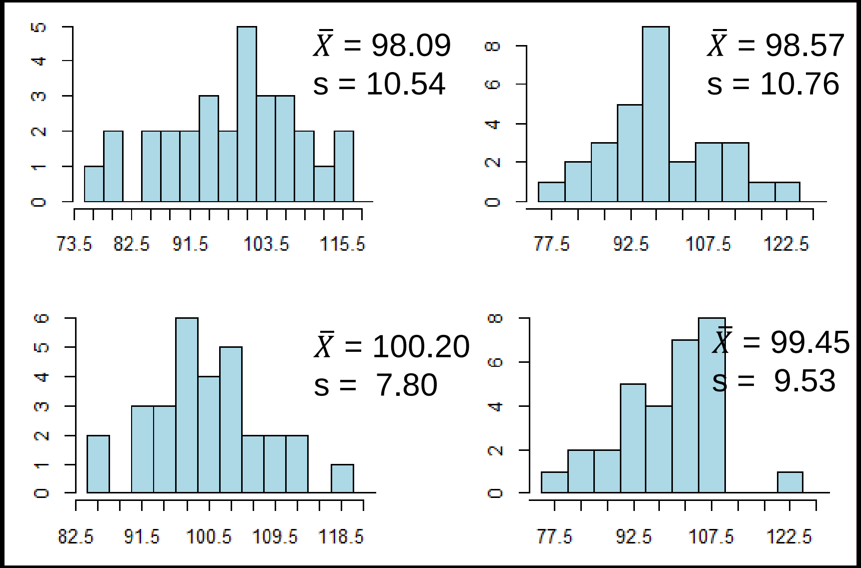
	Sample	**Population**
Definitions	Subgroup or portion of the population chosen for evaluation or study	Collection of all items produced or considered
Characteristics	Statistics	Parameters
Size	n	N
Mean	$\bar{X}$	μ
Median	$\tilde{X}$	M
Standard Deviation	s	σ
Variance	s^2	σ^2
Skewness	g_3	γ_3
Kurtosis	g_4	γ_4
Proportion	p	π
Rate	$\bar{c}$	λ

Sampling Distributions and Statistical Inference

$\mu = 100$

$\sigma = 10$





Creating Random Numbers in R

In R / Rstudio:

```
> rnorm( )
> rexp( )
> rpois( )
> rbinom( )
```

Sampling Error

- Repeated samples may not be identical
- Descriptive statistics calculated from repeated sampling (with replacement) will not be exactly the same, even though the population is unchanged.

Sampling Error

- This is an expected phenomenon since we are not measuring all of the subjects or units for the entire population.
- Statistical methods allow us to account for sampling error, and make appropriate decisions.

Sampling Error

- In spite of the presence of sampling error, random sampling allow us to use sample statistics as point estimators of population parameters; however

Sampling Error

- Even when unbiased, sample statistics will probably not exactly equal their associated true population parameters.

Sampling Error

- An observed difference between a true parameter value and its associated sample descriptive statistic is caused by sampling error.

Sampling Error Defined

- The expected and quantifiable discrepancy between a population parameter and its associated descriptive statistic due to the sample size employed, and in the case of some descriptive statistics, the variability of the population.

Sampling Error & Probability

- Sampling Error is quantifiable using Random Sampling Distributions (RSDs).

- These distributions, like all probability distributions, are based on the principles of classical probability.

Sources

- Luftig, J. An Introduction to Statistical Process Control & Capability. Luftig & Associates, Inc. Farmington Hills, MI, 1982

Random Sampling Distributions

Data Science for Quality Management: Sampling Distributions, Error and Estimation

with **Wendy Martin**

Learning objectives:

Explore the concept of random sampling distributions in R

Random Sampling Distributions

A RSD is the distribution of a sample statistic calculated from all possible random samples of the given (fixed) size from a given population.

Random Sampling Distributions

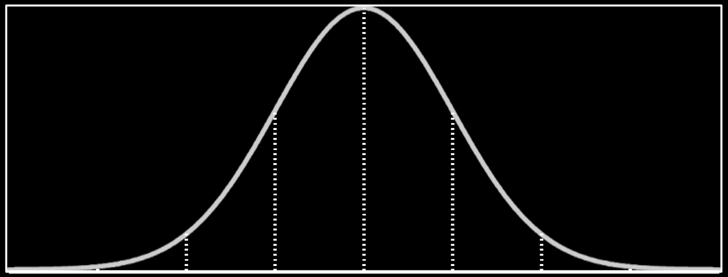
A random sampling distribution is a population distribution and foundational for understanding statistical inference.

Random Sampling Distributions

- Draw all possible random samples of size n from a given research population
- Calculate descriptive statistics for each of the samples
- Construct a distribution for each of the sampled descriptive statistics

Random Sampling Distributions

- Each of the resultant distributions constitutes the random sampling distribution of the statistics.

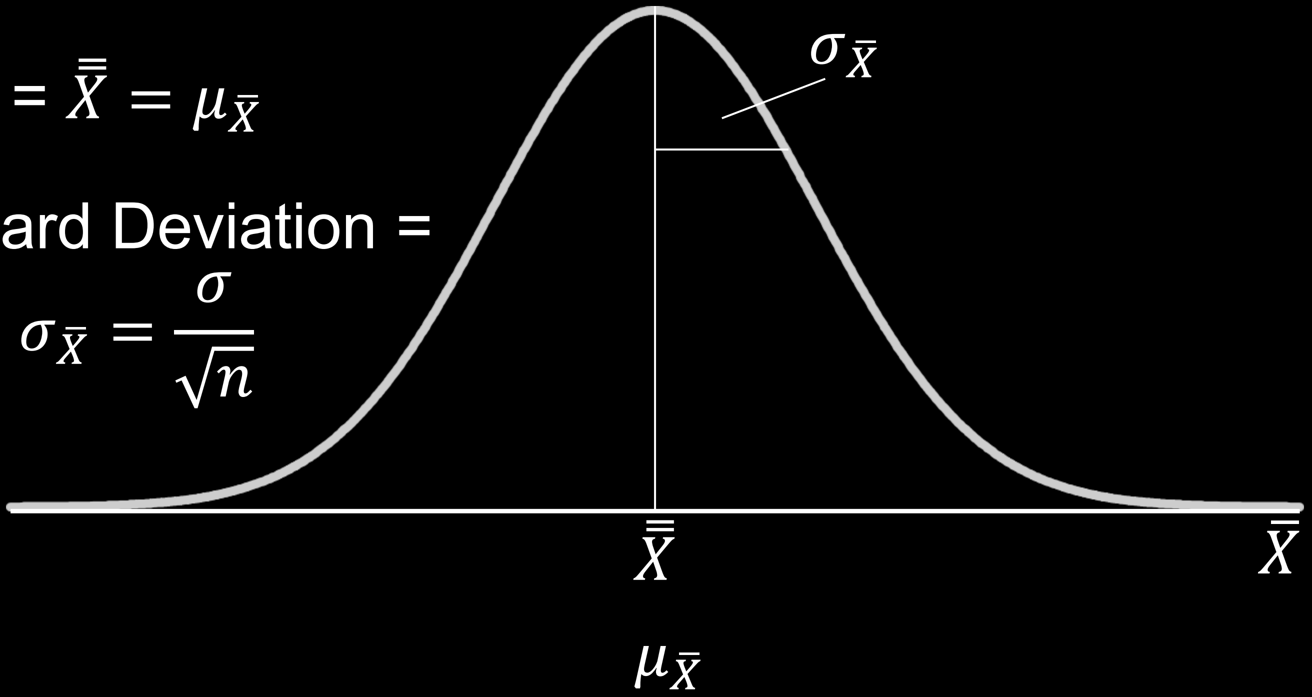


RSD of the Sample Averages

Mean = $\bar{\bar{X}} = \mu_{\bar{X}}$

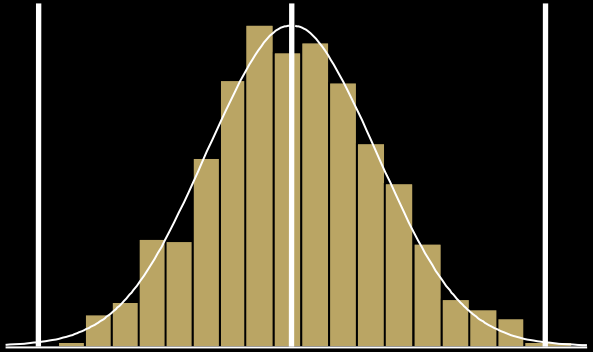
Standard Deviation =

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

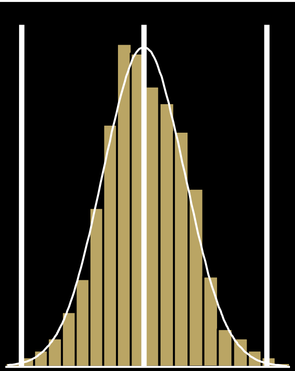


RSD of the Sample Averages

(from a Normally Distributed Population)



Distribution of Individuals



Distribution of Means (n = 4)

Sources

- Luftig, J. An Introduction to Statistical Process Control & Capability. Luftig & Associates, Inc. Farmington Hills, MI, 1982

The Central Limit Theorem

Data Science for Quality Management: Sampling Distributions, Error and Estimation

with **Wendy Martin**

Learning objective:

Describe the Central Limit Theorem

The Central Limit Theorem

- The mean of the RSD of means will equal the population mean, µ, even if the parent population is not normally distributed.

The Central Limit Theorem

- As the sample size (n) increases, the RSD of the means will approach normality, regardless of the shape of the process distribution.

The Central Limit Theorem

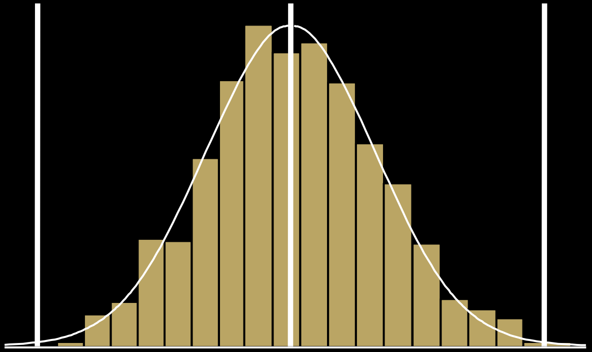
- Based upon this theorem, we can use sample statistics to make inferences about population parameters.

The Central Limit Theorem

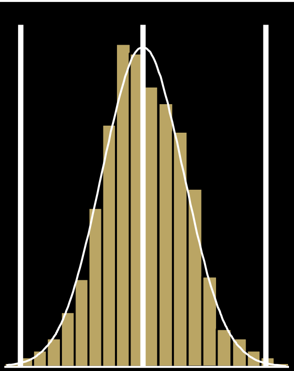
- This applies even without our knowing anything about the shape of that population other than what we can gather from the sample (in most cases).

RSD of the Means

(from a Normally Distributed Population)



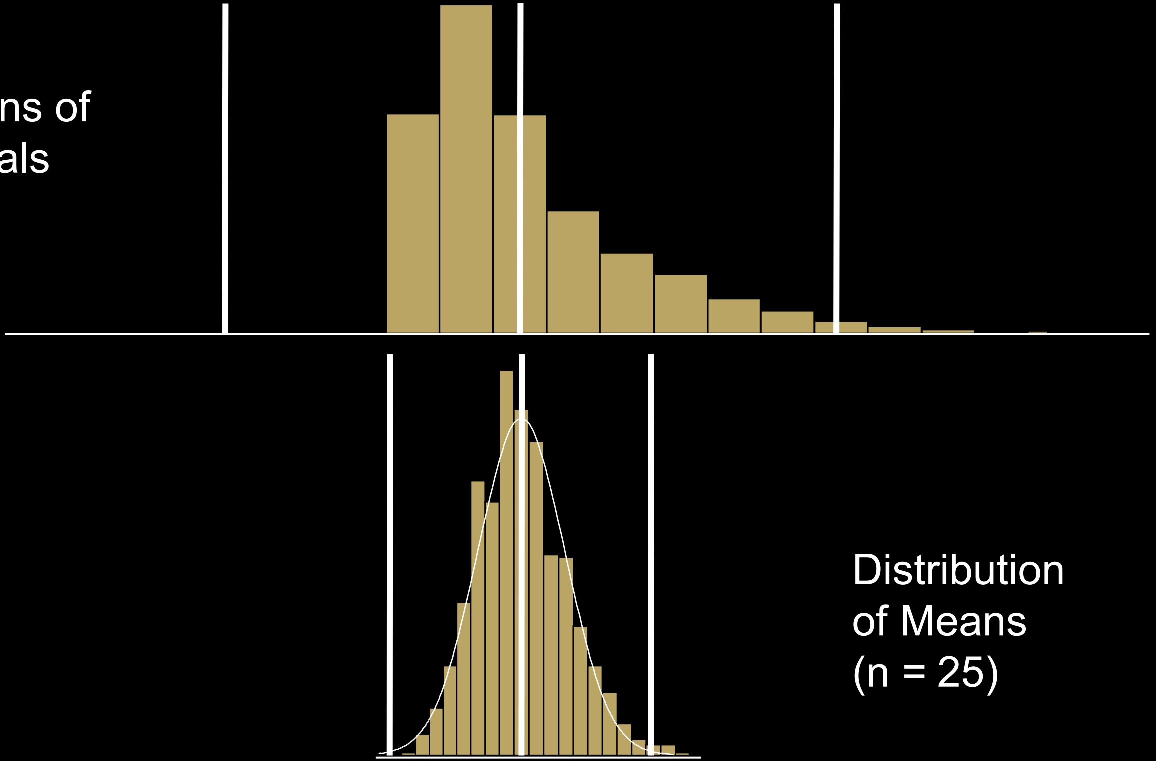
Distribution of Individuals



Distribution of Means (n = 4)

RSD of the Means

(from an Exponentially Distributed Population)



Sources

- Luftig, J. An Introduction to Statistical Process Control & Capability. Luftig & Associates, Inc. Farmington Hills, MI, 1982

Probability with RSDs

Data Science for Quality Management: Sampling Distributions, Error and Estimation

with **Wendy Martin**

Learning objective:

Estimate probability using the Random Sampling Distribution of the mean

Estimating Probability Using the RSD of the Mean

- Note that when we use the Standard Error of the Estimate to find areas on the RSD of the means, the z-score employed becomes:

Estimating Probability Using the RSD of the Mean

$$Z_{\bar{X}} = \frac{\bar{X} - \mu}{\sigma_{\bar{X}}} = \frac{\bar{X} - \mu}{\sigma / \sqrt{n}}$$

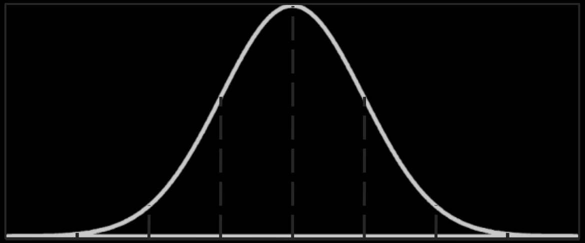
RSD of the Mean – Example 1

- A process has historically manufactured parts at a mean, μ, of 1.325, with a standard deviation, σ, of 0.045.

RSD of the Mean – Example 1

- Drawing a random sample of 25 units, what is the probability of finding an $\bar{X}$ of 1.433 or more for the sample if no change has occurred in the mean or dispersion of the process?

RSD of the Mean – Example 1



$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} = \frac{0.045}{\sqrt{25}} = 0.009$$

$$Z_{\bar{X}} = \frac{\bar{X} - \mu}{\sigma_{\bar{X}}} = \frac{1.433 - 1.325}{0.009} = 12$$

RSD of the Mean – Example 1

- The probability associated with this Z score is….really, really small!

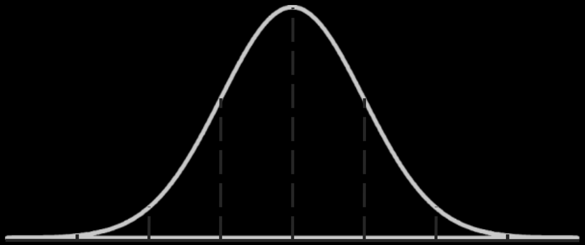
RSD of the Mean – Example 2

- A process has historically manufactured parts at a mean, μ, of 50, with a standard deviation, σ, of 14.4.

RSD of the Mean – Example 2

- Drawing a random sample of 16 units, what is the probability of finding an $\bar{X}$ of 55 or more for the sample if no change has occurred in the mean or dispersion of the process?

RSD of the Mean – Example 2



$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} = \frac{14.4}{\sqrt{16}} = 3.6$$

$$Z_{\bar{X}} = \frac{\bar{X} - \mu}{\sigma_{\bar{X}}} = \frac{55 - 50}{3.6} = 1.389$$

RSD of the Mean – Example 2

- The probability associated with this Z score = 0.0824

RSD Examples

Statistic	RSD	Standard Error
$\bar{X}$	RSD of the mean	of the mean
$\tilde{X}$	RSD of the median	of the median
p	RSD of the proportion	of the proportion
R	RSD of the range	of the range

Sources

- Luftig, J. An Introduction to Statistical Process Control & Capability. Luftig & Associates, Inc. Farmington Hills, MI, 1982

Estimates and Estimators

Data Science for Quality Management: Sampling Distributions, Error and Estimation

with **Wendy Martin**

Learning objectives:

Describe types of estimates

List the criteria for “good” estimators

Types of Estimates

Point Estimate

- A single number used to estimate an unknown parameter

Types of Estimates

Interval Estimate

- A range of values used to estimate a population parameter

Types of Estimates

Estimator

- A sample statistic used to estimate a population parameter. An estimate is a specific observed value of a statistic.

Criteria for “Good” Estimators

- Unbiased
- Efficient
- Consistent
- Sufficient

Criteria for “Good” Estimators

Unbiasedness

- The mean of the Random Sampling Distribution (RSD) of the estimator is equal to the parameter it estimates.

Criteria for “Good” Estimators

Efficiency

- Refers to the standard error of the statistic RSD. The most efficient estimator is the one with the smallest standard error.

Criteria for “Good” Estimators

Consistency

- Refers to the assumption that as n increases, the value of the statistic approaches the value of its associated population parameter.

Criteria for “Good” Estimators

Sufficiency

- Refers to using all possible information in the sample to estimate the corresponding parameter.

Point Estimates

Point Estimate		Population Parameter	
Sample Mean	$\bar{X}$	Population Mean	μ
Sample Variance	s^2	Population Variance	σ^2
Sample Proportion	p	Population Proportion	π
Sample Count	c	Population Count	λ
Sample Skewness	g_3	Population Skewness	γ_3
Sample Kurtosis	g_4	Population Kurtosis	γ_4

Point Estimates

$$\bar{X} \approx \mu$$

$$s \approx \sigma \; and \; s^2 \approx \sigma^2$$

$$p \approx \pi$$

$$c \approx \lambda$$

Estimating σ from Multiple Samples

Average Range

$$\hat{\sigma} = \frac{\bar{R}}{d_2}$$

Median Range

$$\hat{\sigma} = \frac{\tilde{R}}{\tilde{d}_2}$$

Average Standard Deviation

$$\hat{\sigma} = \frac{\bar{s}}{c_4}$$

Median Standard Deviation

$$\hat{\sigma} = \sqrt{\frac{(\tilde{s})^2}{\chi_{0.5,n-1df}}}$$

Estimating σ from Multiple Samples

Average Variance (equal sample size)

$$\hat{\sigma} = \sqrt{\overline{s^2}}$$

(unequal sample size)

$$\hat{\sigma} = \sqrt{\frac{\sum_{j=1}^{k}(n_j - 1)s^2}{\sum_{j=1}^{k}(n_j - 1)}}$$

Estimating σ from Multiple Samples

Average Moving Range of the Mean

$$\hat{\sigma} = \frac{\overline{MR}_{\bar{X}}\sqrt{n}}{d_2}$$

Median Moving Range of the Means

$$\hat{\sigma} = \frac{\widetilde{MR}_{\bar{X}}\sqrt{n}}{\tilde{d}_2}$$

Standard Deviation of the Means $\hat{\sigma} = s_{\bar{X}}\sqrt{n}$

Point Estimates in RStudio

Point Estimate	In RStudio
$\bar{X}$	`mean()`
s	`sd()`
p	`mean() # average proportion`
c	`mean() # average count per unit`

Sources

- Luftig, J. An Introduction to Statistical Process Control & Capability. Luftig & Associates, Inc. Farmington Hills, MI, 1982

Confidence Levels and Interval Estimates

Data Science for Quality Management: Sampling Distributions, Error and Estimation

with **Wendy Martin**

Learning objective:

Differentiate between confidence level and confidence interval

Confidence Level

- The confidence level is the probability associated with an interval estimate.
- This refers to the probability that the interval estimate includes the population parameter.

Confidence Level

- Typical confidence levels used are 90, 95, and 99%, with 95% confidence levels used most frequently
- Alpha, α, is one minus the confidence level

Confidence Interval

- The confidence interval is the range of the estimate. The confidence interval is often expressed in standard error values

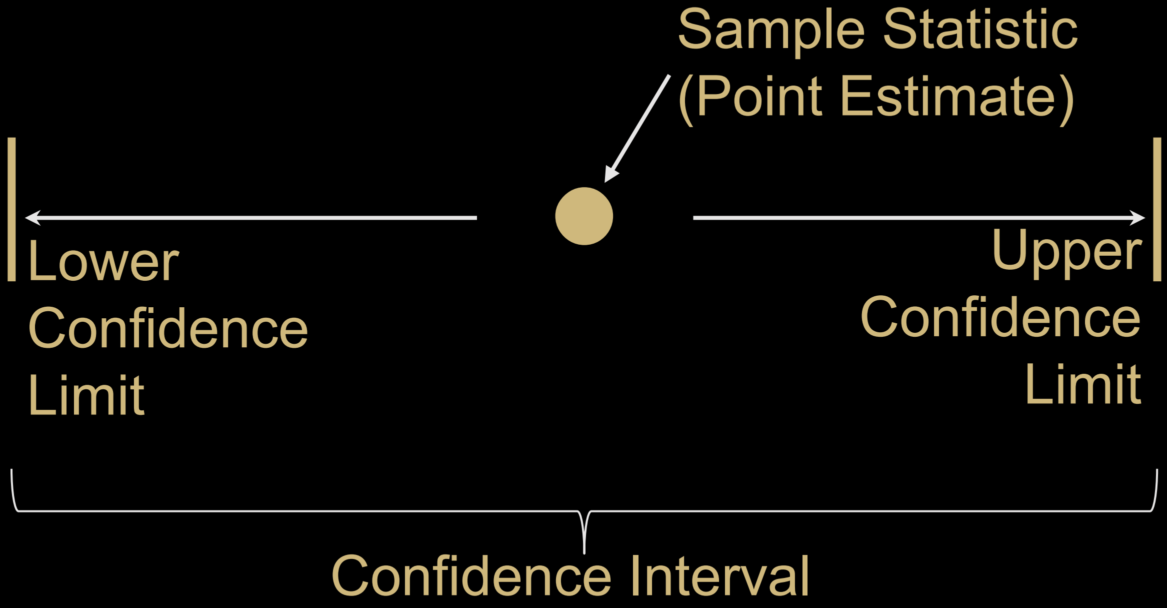
Confidence Intervals

- Confidence intervals provide a range of values in which we would expect to find the true population parameter, with a given level of confidence

Confidence Interval

- A 95% confidence interval for a population mean is the interval that has a 95% probability of the true population mean being found within it

Interval Estimation



Sample Statistic
(Point Estimate)

Lower
Confidence
Limit

Upper
Confidence
Limit

Confidence Interval

Confidence Interval

- An interval estimate provides us a way to qualify our estimate by indicating the magnitude of the sampling error, and hence, the precision of our estimate

Confidence Interval

- To find this interval, we must look at the set of all possible parameters and assess each of those parameters for their probability of providing us with the sample statistic we observed

Interval Estimate Example

- A warranty group wishes to determine the mean life of batteries placed in new cars.
- A sample of 200 batteries is drawn and the mean battery life is found to be 38 months, with a standard deviation of 4 months.

Interval Estimate Example

- Their point estimate for the population mean is 38 months, but they also realize that sampling error is present, and they wish to quantify the uncertainty of this estimate.

Interval Estimate Example

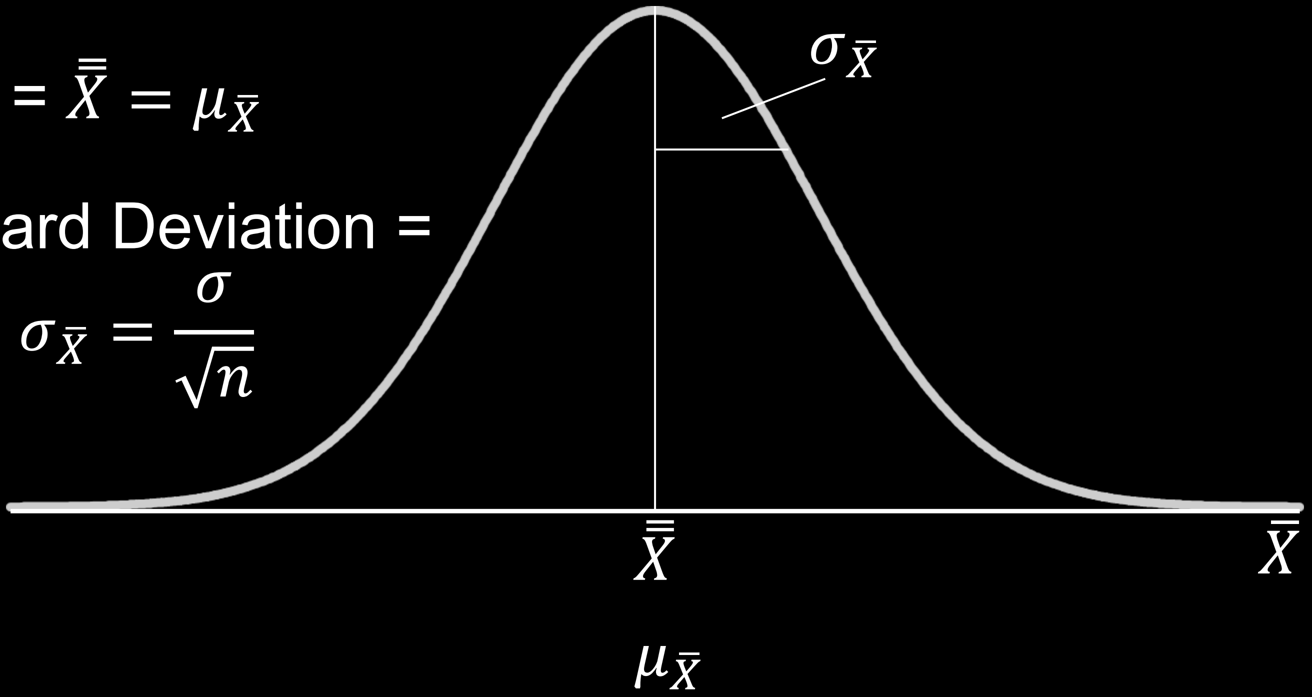
- The question they must answer is what population means could have given a sample mean of 38?

RSD of the Sample Averages

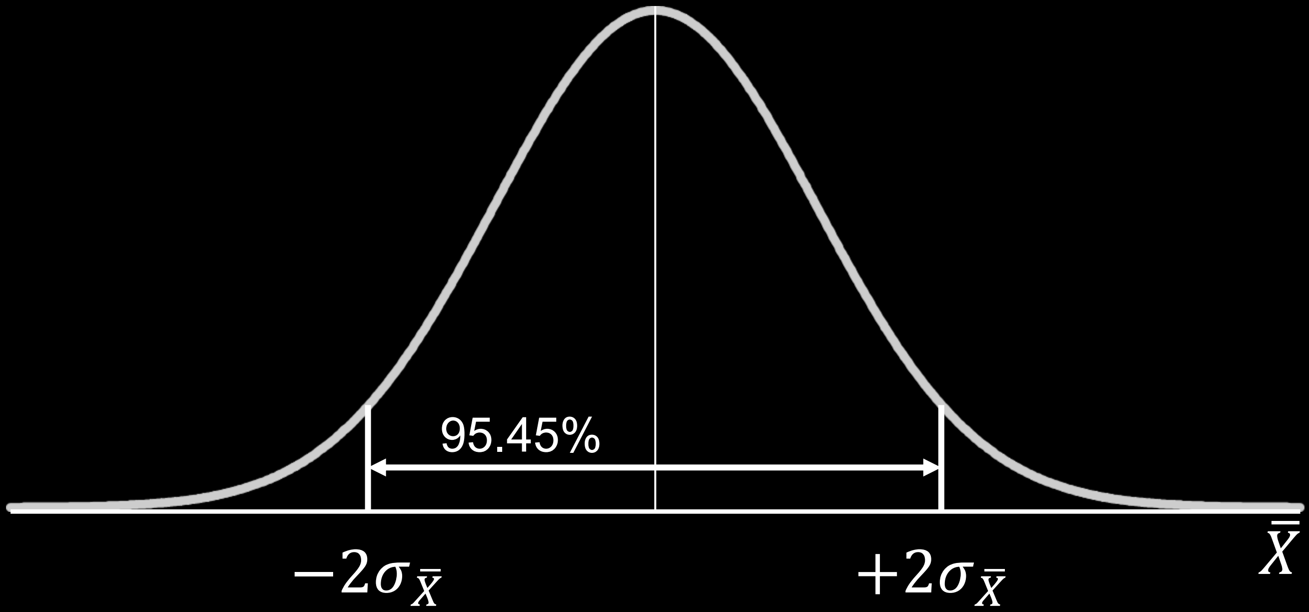
Mean = $\bar{\bar{X}} = \mu_{\bar{X}}$

Standard Deviation =

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



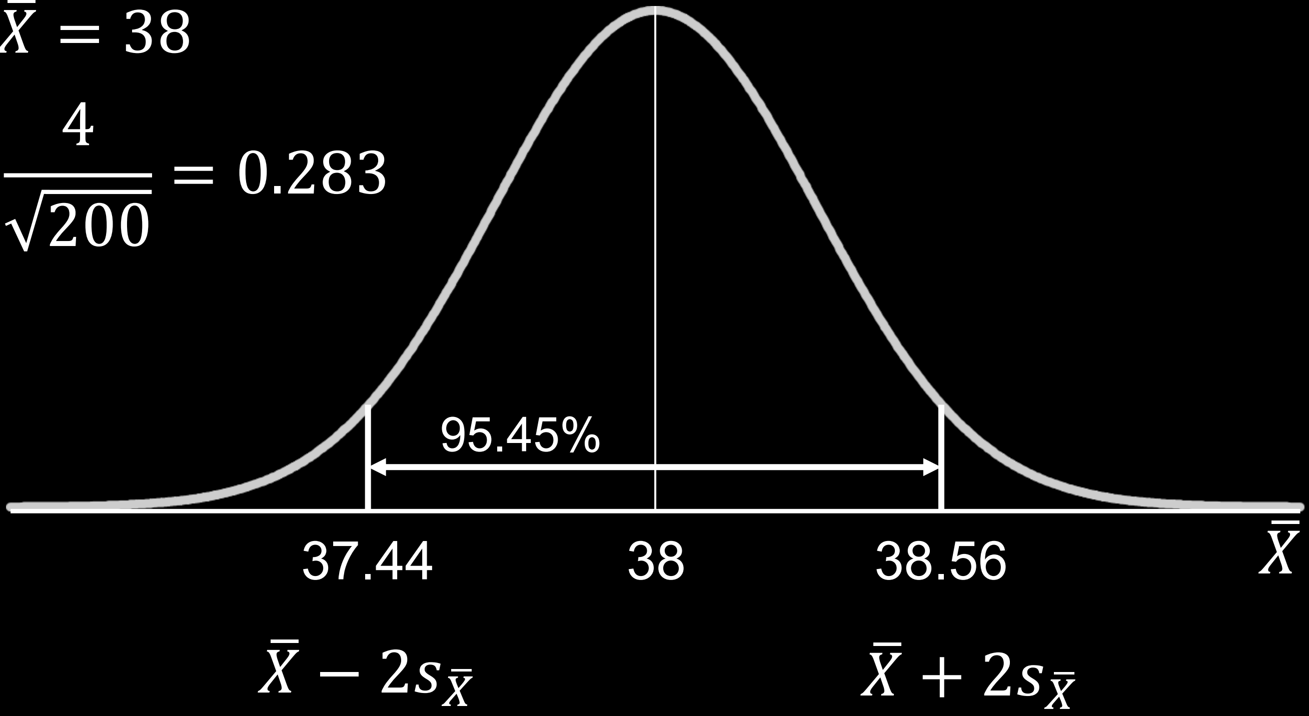
Interval Estimate Example



Interval Estimate Example

$$\mu \approx \bar{X} = 38$$

$$s_{\bar{X}} = \frac{4}{\sqrt{200}} = 0.283$$



Sources

- Luftig, J. An Introduction to Statistical Process Control & Capability. Luftig & Associates, Inc. Farmington Hills, MI, 1982

Confidence Intervals for the Mean and Variance

Data Science for Quality Management: Sampling Distributions, Error and Estimation

with **Wendy Martin**

Learning objective:

Calculate interval estimates for the mean, variance and standard deviation

Calculating Confidence Intervals

- Confidence intervals may be calculated for various statistics
  - Mean
    - Sigma known
    - Sigma unknown
  - Standard Deviation / Variance

Means (Sigma Known)

- If the standard deviation is known, the following formula may be used

$$\mu_{CI} = \bar{X} \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$$

Example

- For example, assume a sample was taken with the following characteristics

$$n = 150,\ \bar{X} = 20,\ s = 5$$

Confidence Level Desired = 95%

Example

- The 95% confidence interval (CI) for the mean is 19.2 to 20.8

$$\mu_{CI} = 20 \pm 1.96\frac{5}{\sqrt{150}} = 20 \pm 0.80$$

Interval Estimate for the Mean (when σ is known)

In RStudio

```
> z.test.onesample
> z.test.onesample.simple
```

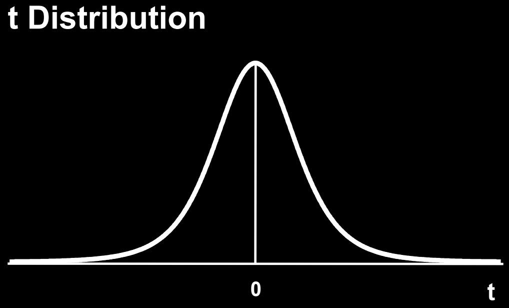
Means (Sigma Unknown)

- When the standard deviation is unknown, there is also uncertainty in the estimate of the standard error

- To account for this, instead of using the *z* distribution, we must use the *t* distribution

The *t* Distribution

- This distribution is commonly employed with small samples and an unknown population *σ*
- The *t* distribution is lower at the mean and higher at the tails than the normal distribution



The *t* Distribution

- The shape of the *t* distribution is dependent upon υ, or (*df*), the degrees of freedom. Degrees of freedom relate to “the number of values which can be freely chosen”

The *t* Distribution

- The *t* distribution considers the fact there is error associated with the use of *s*, the sample standard deviation to estimate the population (σ)

- This error increases the variability of the resulting statistic, the *t*, relative to a standard normal distribution

Means (Sigma Unknown)

- If the standard deviation is unknown, the following formula may be used

$$\mu_{CI} = \bar{X} \pm t_{\alpha/2,(n-1)df} \frac{s}{\sqrt{n}}$$

Example

- For example, assume a sample was taken with the following characteristics

$$n = 14, \bar{X} = 15{,}000, \quad s = 500$$

Confidence Level desired = 90%

Example

- The df are equal to n - 1 or 13
- The t value corresponding to this is 1.771

$$\mu_{CI} = 15000 \pm 1.771 \frac{500}{\sqrt{14}} = 15000 \pm 236.66$$

- The 90% (CI) for the mean is 14763.35 to 15236.65

Interval Estimate for the Mean (when σ is unknown)

In RStudio

```
> t.test.onesample
> t.test.onesample.simple
```

Standard Deviation / Variance

- The following formula may be used to generate a confidence interval for a standard deviation

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

Standard Deviation / Variance

- This formula assumes that the population sampled from may be approximated by the normal distribution

Example

- For example, a process is studied for variability and a sample is drawn with the following characteristics

s = 10 and n = 25

Confidence Level desired = 95%

Example

- The 95% CI for the standard deviation is 7.81 to 13.91

$$\sqrt{\frac{10^2(24)}{39.364}} < \sigma < \sqrt{\frac{10^2(24)}{12.401}}$$

Interval Estimate for the Variance / Standard Deviation

In RStudio

```
> variance.test.onesample
> variance.test.onesample.simple
```

Sources

- Luftig, J. An Introduction to Statistical Process Control & Capability. Luftig & Associates, Inc. Farmington Hills, MI, 1982

Confidence Intervals for Proportions and Poisson Counts

Data Science for Quality Management: Sampling Distributions, Error and Estimation

with **Wendy Martin**

Learning objective:

Calculate interval estimates for proportions and Poisson counts

Calculating Confidence Intervals

- Confidence intervals may be calculated for various statistics
  - Proportion
  - Poisson Counts

Proportion (Exact Binomial)

- Many formulas exist to generate confidence intervals for proportions

- lolcat uses the formula is that based on the exact binomial distribution

Proportion (Exact Binomial)

- The exact binomial confidence interval for a proportion uses quantiles from the Beta distribution

$$\pi_{Lower} = \beta(\frac{\alpha}{2}; np, n - np + 1)$$

$$\pi_{Upper} = \beta(1 - \frac{\alpha}{2}; np + 1, n - np)$$

Example

- For example, assume a sample was taken with the following characteristics

n = 100, p = 0.12
Confidence Level Desired = 95%

Example

- np = 12
- n = 100
- α = 0.05

$$\pi_{Lower} = qbeta(0.025; 12{,}89) = 0.0636$$

$$\pi_{Upper} = qbeta(0.975; 13{,}88) = 0.2002$$

Interval Estimate for a Proportion

In RStudio

```
> proportion.test.onesample.exact
> proportion.test.onesample.exact.simple
```

Poisson Counts

- Even more formulas exist to generate confidence intervals for Poisson Counts
- lolcat uses the formula is that based on the exact Poisson distribution

Poisson Counts

- The exact Poisson confidence interval for a Poisson Count uses the quantile values from the Gamma distribution

$$\lambda_{Lower} = \frac{G(\frac{\alpha}{2}, x)}{n} \qquad \lambda_{Upper} = \frac{G(1 - \frac{\alpha}{2}, x + 1)}{n}$$

where x = λ*n

Example

- For example, assume a sample was taken with the following characteristics

$n = 20, \lambda = 25.05$
Confidence Level Desired = 95%

Example

- λ = 25.05
- n = 20
- $\lambda * \text{n} = 501$
- α = 0.05

$$\lambda_{Lower} = qgamma(0.025; 501)/20 = 22.90$$

$$\lambda_{Upper} = qgamma(0.975; 502)/20 = 27.34$$

Example

- λ = 25.05
- n = 20
- $\lambda * \mathrm{n} = 501$
- α = 0.05

$$\lambda_{Lower} = qgamma(0.025; 501)/20 = 23.24$$

$$\lambda_{Upper} = qgamma(0.975; 502)/20 = 26.97$$

Interval Estimate for Poisson Count

In RStudio

```
> poisson.test.onesample.simple()
```

Sources

- Luftig, J. An Introduction to Statistical Process Control & Capability. Luftig & Associates, Inc. Farmington Hills, MI, 1982