Parameters estimation

Outline

Point estimate and construct confidence interval for popular parameters

- mean population μ
- variance population σ^2
- proportion population p

Based on sample data, one makes inferences or generalizations about population parameter

Statistics inference: generalize and prediction

State that the average cost to build a residence in Charleston, South Carolina, is between \$330,000 and \$335,000, based on the estimates of 3 contractors selected at random from the 30 now building in this city

Inference about Population from Sample Information

- compute statistics from a selected sample from a population
- **From this statistics**, make some **statement about a parameter** of a population

Two Majors in Statistic inference

- Estimation
- Hypothesis testing

Estimation

- Population parameters (unknown): Mean, variance, standard deviation ...
- Statistics (from data): Sample mean, sample variance ...
- Use statistics to estimate parameter: point estimate
- How accurate: interval estimate

A random sample of size n is a sequence of RVs

$$X_1, \ldots, X_n$$

Point estimate

A point estimate of some population parameter θ from random sample $X_1, X_2, \ldots, X_n$ is a single value $\hat{\theta}$ of a statistic $\hat{\Theta} = \hat{\Theta}(X_1, \ldots, X_n)$

Example

- A value of sample mean

$$\bar{X} = \frac{X_1 + \ldots X_n}{n}$$

  is a point estimate of the population mean μ.

- $\hat{p} = \frac{x}{n}$ is a point estimate of the true proportion p for a binomial experiment.

Example

- value of sample mean depends on the sample that you observe
- sample is chosen randomly $\rightarrow$ different value of sample mean for difference sample
- Sample mean is a Random variable

- We do not expect $\bar{X}$ to estimate μ exactly, but we certainly hope that it is not far off.
- it is possible to obtain a closer estimate of μ by using the sample median $\tilde{X}$ as an estimator
- Not knowing the true value of μ, we must decide in advance whether to use $\bar{X}$ or $\tilde{X}$ as our estimator.
- **What are the desirable properties of a "good" decision function that would influence us to choose one estimator rather than another?**

- We do not expect $\bar{X}$ to estimate μ exactly, but we certainly hope that it is not far off.
- it is possible to obtain a closer estimate of μ by using the sample median $\tilde{X}$ as an estimator
- Not knowing the true value of μ, we must decide in advance whether to use $\bar{X}$ or $\tilde{X}$ as our estimator.
- **What are the desirable properties of a "good" decision function that would influence us to choose one estimator rather than another?**

- We do not expect $\bar{X}$ to estimate μ exactly, but we certainly hope that it is not far off.
- it is possible to obtain a closer estimate of μ by using the sample median $\tilde{X}$ as an estimator
- Not knowing the true value of μ, we must decide in advance whether to use $\bar{X}$ or $\tilde{X}$ as our estimator.
- What are the **desirable properties of a "good"** decision function that would influence us to choose one estimator rather than another?

- We do not expect $\bar{X}$ to estimate μ exactly, but we certainly hope that it is not far off.
- it is possible to obtain a closer estimate of μ by using the sample median $\tilde{X}$ as an estimator
- Not knowing the true value of μ, we must decide in advance whether to use $\bar{X}$ or $\tilde{X}$ as our estimator.
- **What are the desirable properties of a "good" decision function that would influence us to choose one estimator rather than another?**

Unbiased estimator

A statistics $\hat{\Theta}(X_1, \ldots, X_n)$ is said to be an unbiased estimator for (population) parameter θ if

$$E(\hat{\Theta}) = \theta$$

Example

Population with mean μ and variance σ^2

- Sample mean $\bar{X} = \frac{X_1 + \ldots X_n}{n}$

$$E(\bar{X}) = \mu$$

- Sample variance $S^2 = \frac{(X_1 - \bar{X})^2 + \ldots (X_n - \bar{X})^2}{n-1}$

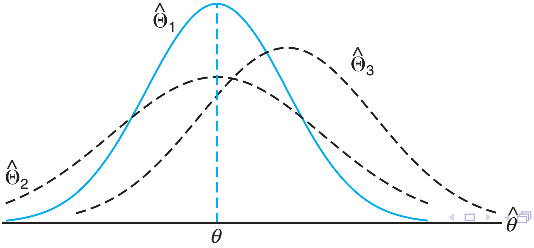
$$E(S^2) = \sigma^2$$

Sample mean $\bar{X}$ and sample variance S^2 are unbiased estimators of μ and σ^2 respectively

Variance of estimator

- $\hat{\Theta}_1, \hat{\Theta}_2$: unbiased estimators for θ
- $\hat{\Theta}_1$ is a more efficient estimator than $\hat{\Theta}_2$ if

$$Var(\hat{\Theta}_1) \leq Var(\hat{\Theta}_2)$$



Efficient estimator

- The most efficient estimator: unbiased estimator with smalles variance
- $\bar{X}$ and S^2 are the most efficient estimators of μ and σ^2

Example

Given the sample data

$$1, \quad 1, \quad 4, \quad 6$$

Find the best point esimators for the population mean and population variance

Solution

The best estimator for the population mean is the sample mean

$$\bar{x} = \frac{1+1+4+6}{4} = 3$$

The best esimator for the population variance is the sample variance

$$s^2 = \frac{(1-3)^2+(1-3)^2+(4-3)^2+(6-3)^2}{4-1} = 6$$

Solution

The best estimator for the population mean is the sample mean

$$\bar{x} = \frac{1+1+4+6}{4} = 3$$

The best esimator for the population variance is the sample variance

$$s^2 = \frac{(1-3)^2 + (1-3)^2 + (4-3)^2 + (6-3)^2}{4-1} = 6$$

Interval estimate

- estimation accuracy increases with large samples
- but don't expect $\bar{X}$ to be exactly μ
- Want to find an interval around $\bar{X}$ so we can be sure that μ is in it.
- Ex: want to find $[a, b]$ so that 95% of the time $\mu \in [a, b]$
- $[a, b]$ is called *95% confidence interval estimate* of μ

Interval Estimates

An interval estimate of a population parameter θ is an interval of the form

$$\hat{\theta}_L < \theta < \hat{\theta}_U$$

where $\hat{\theta}_L$ and $\hat{\theta}_U$ depend on the value of the statistic $\hat{\theta}$ for a particular sample and also on the distribution of $\hat{\theta}$

- different samples will generally yield different values of $\hat{\theta}$ and different values for $\hat{\theta}_L$ and $\hat{\theta}_U$
- These end points $\hat{\theta}_L$ and $\hat{\theta}_U$ are random variables
- If distribution of $\hat{\theta}$ is known then we can determine

$$P(\hat{\theta}_L < \theta < \hat{\theta}_U) = 1 - \alpha$$

then we have a probability of $1 - \alpha$ of selecting a random sample that will produce an interval containing θ

- different samples will generally yield different values of $\hat{\theta}$ and different values for $\hat{\theta}_L$ and $\hat{\theta}_U$
- These end points $\hat{\theta}_L$ and $\hat{\theta}_U$ are random variables
- If distribution of $\hat{\theta}$ is known then we can determine

$$P(\hat{\theta}_L < \theta < \hat{\theta}_U) = 1 - \alpha$$

then we have a probability of $1 - \alpha$ of selecting a random sample that will produce an interval containing θ

- different samples will generally yield different values of $\hat{\theta}$ and different values for $\hat{\theta}_L$ and $\hat{\theta}_U$
- These end points $\hat{\theta}_L$ and $\hat{\theta}_U$ are random variables
- If distribution of $\hat{\theta}$ is known then we can determine

$$P(\hat{\theta}_L < \theta < \hat{\theta}_U) = 1 - \alpha$$

then we have a probability of $1 - \alpha$ of selecting a random sample that will produce an interval containing θ

- The interval

$$\hat{\theta}_L < \theta < \hat{\theta}_U$$

computed from the selected sample is called a $100(1-\alpha)\%$ **confident interval**

- The fraction $100(1-\alpha)\%$: **confidence coefficient** or **degree of confidence**

- $\hat{\theta}_L$ and $\hat{\theta}_U$: lower and upper **confident limits**

- The interval

$$\hat{\theta}_L < \theta < \hat{\theta}_U$$

computed from the selected sample is called a $100(1-\alpha)\%$ **confident interval**

- The fraction $100(1-\alpha)\%$: **confidence coefficient** or **degree of confidence**

- $\hat{\theta}_L$ and $\hat{\theta}_U$: lower and upper **confident limits**

- The interval
$$\hat{\theta}_L < \theta < \hat{\theta}_U$$
computed from the selected sample is called a $100(1-\alpha)\%$ **confident interval**
- The fraction $100(1-\alpha)\%$: **confidence coefficient** or **degree of confidence**
- $\hat{\theta}_L$ and $\hat{\theta}_U$: lower and upper **confident limits**

Estimate the mean when the variance population σ^2 is known

Select a random sample of size n, $X_1, \ldots, X_n$, from a population with mean μ and finite variance σ^2.

Sample mean

$$\bar{X} = \frac{X_1 + \ldots + X_n}{n}$$

is used to estimate the true mean μ of the population - called a point estimate of μ

Properties of sample mean

- Expectation

$$E(\bar{X}) = \mu_{\bar{X}} = \mu$$

unbiased estimator of the true mean μ

- Variance

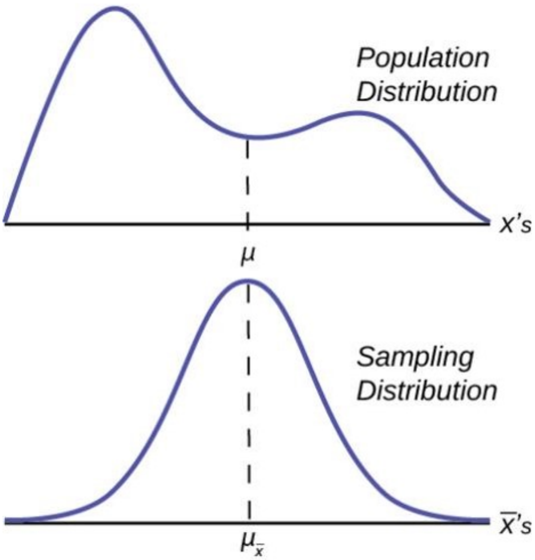
$$Var(\bar{X}) = \sigma^2_{\bar{X}} = \frac{\sigma^2}{n}$$

Sample mean from normal population

- Population has normal distribution $\mathrm{N}(\mu, \sigma^2)$
- Each observation $X_1, \ldots, X_n \sim \mathrm{N}(\mu, \sigma^2)$
- $(X_1 + \cdots + X_n) \sim \mathrm{N}(n\mu, n\sigma^2)$
- Sample mean $\bar{X} \sim \mathrm{N}(\mu, \sigma^2/n)$

Central limit theorem

- Suppose $X_1, \ldots, X_n$ i.i.d with mean μ and variance σ^2.
- then for n large enough, $\bar{X}$ has distribution approximately normal with mean μ and variance $\dfrac{\sigma^2}{n}$.


Population
Distribution
X's
μ
Sampling
Distribution
$\overline{X}$'s
$\mu_{\bar{x}}$

Distribution of $\bar{X}$

- Sample mean

$$\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n} \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right)$$

- Standadize

$$\frac{\bar{X} - \mu}{\sigma / \sqrt{n}} \sim \mathcal{N}(0, 1)$$

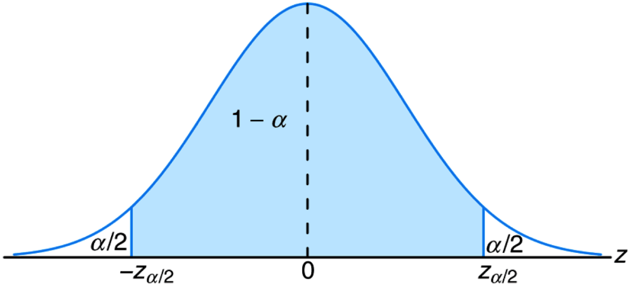
- Valid for large sample ($n \geq 30$) or not severely nonnormal population

For sample size $n \geq 30$ or normal population

$$\frac{\bar{X} - \mu}{\sigma / \sqrt{n}} \sim \mathrm{N}(0, 1)$$

$z_{\alpha/2}$ is the critical value determined by

$$P(Z > z_{\alpha/2}) = \alpha/2$$



$$P(-z_{\alpha}/2 < Z < z_{\alpha/2}) = 1 - \alpha$$

$$P\left(-z_{\alpha/2} < \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} < z_{\alpha/2}\right) = 1-\alpha$$

Hence

$$P\left(\bar{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right) = 1-\alpha$$

a $100(1-\alpha)\%$ confidence interval (CI) for population mean μ

$$\boxed{\bar{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}}$$

(two-sided CI)

$100(1-\alpha)\%$: confidence level

- variance population σ^2 known
- Normal population or large sample size $n \geq 30$

a $100(1-\alpha)\%$ confidence interval (CI) for population mean μ

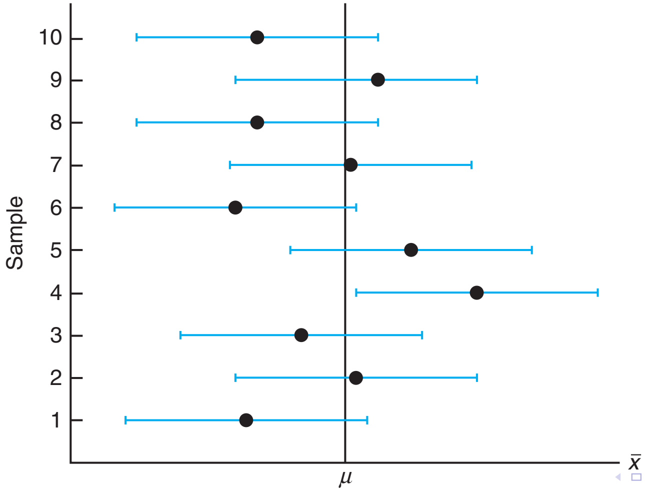
$$\boxed{\bar{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}}$$

(two-sided CI)

$100(1-\alpha)\%$: confidence level

- variance population σ^2 known
- Normal population or large sample size $n \geq 30$

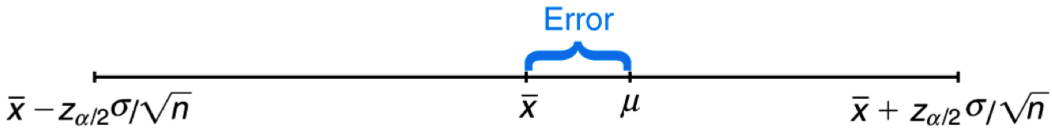
Interval estimate for different samples



Remark

- Given $z_{\frac{\alpha}{2}}$, all of these intervals are of the same width, since their widths once $\bar{x}$ is determined
- The larger the value $z_{\frac{\alpha}{2}}$ is, the wider the intervals are and the more confident we can be that the particular sample selected will produce an interval that contains the unknown parameter μ
- For each $z_{\frac{\alpha}{2}}$, $100(1-\alpha)\%$ of the intervals will cover μ

If $\bar{x}$ is used as an estimate of μ, we can be confident that the error will not exceed $z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$ at $100(1-\alpha)\%$.

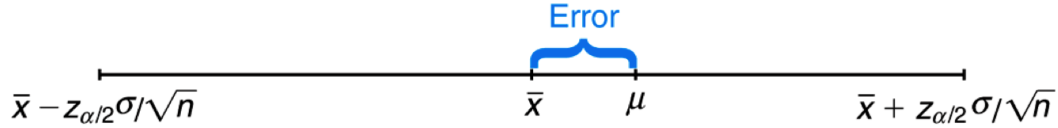


Margin of error $ME = z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$ is used to evaluate accuracy of estimation.

Alternative formular for IC

$$(\bar{x} - ME, \bar{x} + ME)$$

If $\bar{x}$ is used as an estimate of μ, we can be confident that the error will not exceed $z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$ at $100(1-\alpha)\%$.



Margin of error $ME = z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$ is used to evaluate accuracy of estimation.

Alternative formular for IC

$$(\bar{x} - ME, \bar{x} + ME)$$

Example

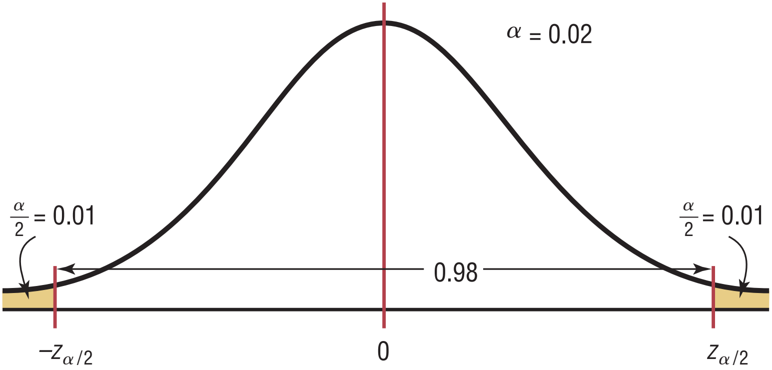
The average zinc concentration recovered from a *sample of measurements taken in 36 different locations* in a river is found to be 2.6 grams per milliliter. **Find the 98% CIs for the mean zinc concentration in the river.** Assume that the population standard deviation is .3 gram per milliliter.

Solution for 98% CI

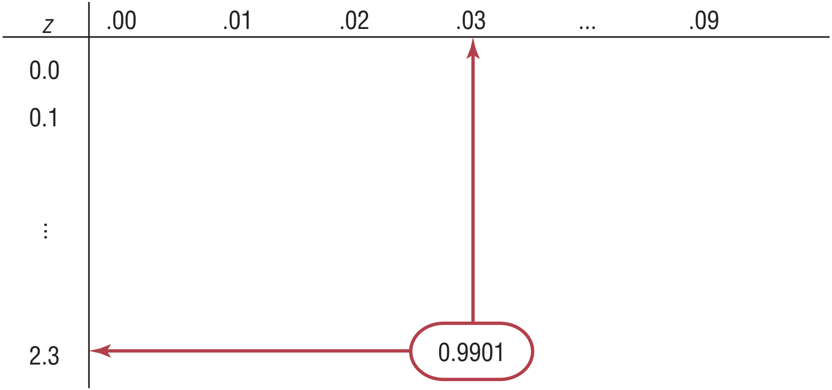
- μ: the mean zinc concentration in the river
- population std $\sigma = .3$
- sample size $n = 36$, sample mean $\bar{x} = 2.6$ g/mil

- Find $\alpha/2$ from level of confidence

  $1 - \alpha = 98\% = .98 \Rightarrow \alpha/2 = .01$



- Find $z_{\frac{\alpha}{2}} = 2.33$



In Excel, NORMINV(1 - $\frac{\alpha}{2}$, 0, 1)

- Marginal of error
  $ME = z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} = 2.33 \times \frac{.3}{\sqrt{36}} = .1165$
- lower bound $\bar{x} - ME = 2.6 - .1165 = 2.4835$
- upper bound $\bar{x} + ME = 2.6 + .1165 = 2.7165$
- 98% CI

$$2.4835 < \mu < 2.7165$$

Practice

A survey of 30 emergency room patients found that the average waiting time for treatment was 174.3 minutes. Assuming that the population standard deviation is 46.5 minutes, find the **best point estimate** of the population mean and the **99% confidence of the** **population mean**

Sample size

Given level of confidence and accuracy ME, one can determine necessary size of sample

$$n = \left(\frac{z_{\frac{\alpha}{2}} \sigma}{ME} \right)^2$$

Example

From past experience it is known that the weights of salmon grown at a commercial hatchery are normal with a mean that varies from season to season but with a standard deviation that remains fixed at 0.3 pounds. If we want to be **95 percent certain** that our **estimate** of the present **season's mean weight of a salmon is correct to within** ± 0.1 **pounds, how large a sample** is needed?

Solution

- Find sample size n such that $ME = .1$
- Information
  - population std $\sigma = .3$
  - Confidence level $1 - \alpha = .95 \Rightarrow z_{\frac{\alpha}{2}} = 1.96$
- $ME = z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} = 1.96 \frac{0.3}{\sqrt{n}} = .1$
- then $n = 35$

An electrical firm manufactures light bulbs that have a length of life that is approximately normally distributed with a standard deviation of 40 hours.

1. If a sample of 30 bulbs has an average life of 780 hours, find a 96% confidence interval for the population mean of all bulbs produced by this firm
2. How large a sample is needed if we wish to be 96% confident that our sample mean will be within 10 hours of the true mean?

Estimate the mean when the population variance σ^2 is not known

In statistics - replace unknown population standard deviation σ by computable sample standard deviation S

$$\frac{\bar{X} - \mu}{S/\sqrt{n}}$$

Distribution of statistics? depends on

- distribution of sample mean $\bar{X}$
- distribution of sample standard deviation S

Estimate the mean when the population variance σ^2 is not known

In statistics - replace unknown population standard deviation σ by computable sample standard deviation S

$$\frac{\bar{X} - \mu}{S/\sqrt{n}}$$

Distribution of statistics? depends on

- distribution of sample mean $\bar{X}$
- distribution of sample standard deviation S

Estimate the mean when the population variance σ^2 is not known

In statistics - replace unknown population standard deviation σ by computable sample standard deviation S

$$\frac{\bar{X} - \mu}{S/\sqrt{n}}$$

Distribution of statistics? depends on

- distribution of sample mean $\bar{X}$
- distribution of sample standard deviation S

$$S^2 = \frac{(X_1 - \bar{X})^2 + \cdots + (X_n - \bar{X})^2}{n-1}$$

Chi-square distribution

- $X_1, \ldots, X_n$ i.i.d. $\mathrm{N}(0,1)$
- $Y = X_1^2 + \cdots + X_n^2$ is said to have the chi-square distribution with n degree of freedom, $Y \sim \chi_n^2$

Distribution of sample variance S^2?

$$S^2 = \frac{(X_1 - \bar{X})^2 + \cdots + (X_n - \bar{X})^2}{n-1}$$

Chi-square distribution

- $X_1, \ldots, X_n$ i.i.d. $\mathrm{N}(0, 1)$
- $Y = X_1^2 + \cdots + X_n^2$ is said to have the chi-square distribution with n degree of freedom, $Y \sim \chi_n^2$

$$S^2 = \frac{(X_1 - \bar{X})^2 + \cdots + (X_n - \bar{X})^2}{n-1}$$

Distribution of sample variance for normal distribution

$$\chi^2 = \frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}$$

Distribution of sample mean $\bar{X}$ for unknown σ^2

Statistics

$$T = \frac{\bar{X} - \mu}{S/\sqrt{n}}$$

t - distribution

If $Z \sim \mathrm{N}(0, 1)$ and $C \sim \chi^2_n$ then $T = \frac{Z}{\sqrt{C/n}}$ has $t-$distribution of n degree of freedom, denoted by $T \sim T(n)$

Statistics

$$T = \frac{\bar{X} - \mu}{S/\sqrt{n}}$$

t - distribution

If $Z \sim \text{ N}(0, 1)$ and $C \sim \chi^2_n$ then $T = \frac{Z}{\sqrt{C/n}}$ has $t-$distribution of n degree of freedom, denoted by $T \sim T(n)$

For normal distribution

- $\bar{X}$: normal distribution
- S^2: related to $\chi^2(n-1)$

Statistics

$$T = \frac{\bar{X} - \mu}{S/\sqrt{n}} \sim t(n-1)$$

For normal distribution

- $\bar{X}$: normal distribution
- S^2: related to $\chi^2(n-1)$

Statistics

$$T = \frac{\bar{X} - \mu}{S/\sqrt{n}} \sim t(n-1)$$

$100(1-\alpha)\%$ CI of μ

- normal population
- variance population σ^2 unknown

$$\boxed{\left(\bar{X} - t_{\alpha/2,n-1}\frac{S}{\sqrt{n}}, \bar{X} + t_{\alpha/2,n-1}\frac{S}{\sqrt{n}}\right)}$$

Margin of error

$$ME = t_{\alpha/2,n-1}\frac{S}{\sqrt{n}}$$

$100(1-\alpha)\%$ CI of μ

- normal population
- variance population σ^2 unknown

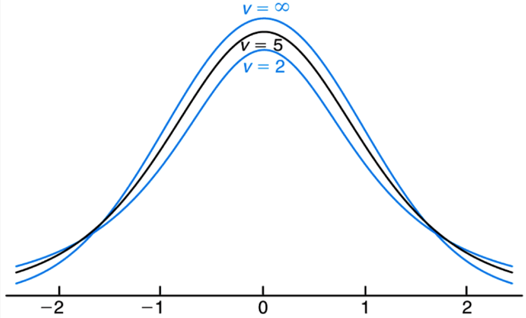
$$\boxed{\left(\bar{X} - t_{\alpha/2,n-1}\frac{S}{\sqrt{n}}, \bar{X} + t_{\alpha/2,n-1}\frac{S}{\sqrt{n}}\right)}$$

Margin of error

$$ME = t_{\alpha/2,n-1}\frac{S}{\sqrt{n}}$$

t - distribution looks like

t is symmetric about 0



degree of freedom $\geq$ 30 then t is approximated by $\mathcal{N}(0, 1)$

Remark

For sample size large $n \geq 30$) then t is approximated by $\mathcal{N}(0, 1)$ so

$$t_{\alpha/2,n-1} \approx z_{\alpha/2}$$

for n large enough

Example

The contents of seven similar containers of sulfuric acid are 9.8, 10.2, 10.4, 9.8, 10.0, 10.2, and 9.6 liters. Find a **95% confidence interval for the mean contents of all such containers**, assuming an approximately normal distribution.

Solution

- μ: mean contents of all containers
- Sample size $n = 7$, sample mean $\bar{x} = 10.0$, sample std $s = 0.283$
- Find $\frac{\alpha}{2}$ from confidence level

$$1 - \alpha = .95 \Rightarrow \frac{\alpha}{2} = .025$$

- Find $t_{\alpha/2,n-1} = t_{.025,6} = 2.447$

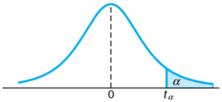


Table A.4 Critical Values of the t-Distribution

v	0.40	0.30	0.20	0.15	0.10	0.05	0.025
1	0.325	0.727	1.376	1.963	3.078	6.314	12.706
2	0.289	0.617	1.061	1.386	1.886	2.920	4.303
3	0.277	0.584	0.978	1.250	1.638	2.353	3.182
4	0.271	0.569	0.941	1.190	1.533	2.132	2.776
5	0.267	0.559	0.920	1.156	1.476	2.015	2.571
6	0.265	0.553	0.906	1.134	1.440	1.943	2.447
7	0.263	0.549	0.896	1.119	1.415	1.895	2.365

- $ME = t_{\alpha/2,n-1}\frac{s}{\sqrt{n}} = 2.447 \times \frac{.283}{\sqrt{7}} = .26$
- Lower bound $\bar{x} - ME = 10.0 - .26 = 9.74$
- Upper bound $\bar{x} + ME = 10.0 + .26 = 10.26$
- 95% CI

$$9.74 \leq \mu \leq 10.26$$

Practice

The following measurements were recorded for the drying time, in hours, of a certain brand of latex paint: 3.4, 2.5, 4.8, 2.9, 3.6, 2.8, 3.3, 5.6, 3.7, 2.8, 4.4, 4.0, 5.2, 3.0, 4.8 Assuming that the measurements represent a random sample from a normal population, find a **95% confidence interval for the average drying time of the paint.**

Remark

For large sample size $n \geq 30$ then

$$t_{\frac{\alpha}{2},n-1} \approx z_{\frac{\alpha}{2}}$$

Hence CI of μ is

$$\boxed{(\bar{X} - z_{\alpha/2}\frac{S}{\sqrt{n}}, \bar{X} + z_{\alpha/2}\frac{S}{\sqrt{n}})}$$

Practice

Scholastic Aptitude Test (SAT) mathematics scores of a random sample of 500 high school seniors in the state of Texas are collected, and the sample mean and standard deviation are found to be 501 and 112, respectively. Find a **99% confidence interval on the mean SAT mathematics score for seniors in the state of Texas.**

Estimate a proportion - mean of Bernoulli RV

- Sample n independent trials from a population, each success with unknown probability p
- Each observation $X_1, \ldots, X_n \sim \text{Ber}(p)$ has two value 1 - success and 0 - failure
- $X = X_1 + \cdots + X_n$: number of successes in n sample trials
- point estimate for p is $\hat{p} = \bar{X} = \frac{X}{n}$ - fraction of successes in sample
- Want to find confidence interval for p

Estimator of population proportion

- Point estimate
$$\hat{p} = \frac{X}{n}$$
then $E(\hat{p}) = p$
- Use $\hat{p}$ as unbiased estimator for p
- For large sample size, by central limit theorem
$$\frac{X - np}{\sqrt{p(1-p)}} = \frac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{n}}} \approx Z \sim N(0, 1)$$

Contruct inteval confidence for p

$$P(-z_{\alpha/2} < Z < z_{\alpha/2}) = 1 - \alpha$$

So

$$P\left(\hat{p} - z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}} < p < \hat{p} + z_{\alpha/2}\underbrace{\sqrt{\frac{p(1-p)}{n}}}_{\text{standard error of point estimator}}\right)$$

However p is unknown. Replace p by $\hat{p}$ in standard error term

$100(1-\alpha)\%$ IC for p

$$\boxed{\hat{p} - z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} < p < \hat{p} + z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}}$$

with

$$ME = z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

Sample size

$$n = (\hat{p})(1-\hat{p})\left(\frac{z_{\alpha/2}}{ME}\right)^2$$

Example

In a random sample of 85 automobile engine crankshaft bearings, 10 have a surface finish that is rougher than the specifications allow.

1. Find a **95% two-sided confidence interval for** p **- the proportion of bearings in the population that exceeds the roughness specification**.

2. **How large a sample** is required if we want to be 95% confident that the error in using $\hat{p}$ to estimate p is less than 0.05

1

- Information
  - sample size $n = 85$
  - sample proportion of bearing that exceeds... : $\hat{p} = \frac{x}{n} = \frac{10}{85} \approx 0.12$
  - Confidence level $1 - \alpha = .95 \Rightarrow z_{\alpha/2} = 1.96$
- $ME = z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = 1.96\sqrt{\frac{(.12)(.88)}{85}} \approx .07$
- 95%CI for p

$$.12 - .07 < p < .12 + .07 \text{ or } .05 < p < .19$$

2. Need to find sample size n such that

$$ME = 0.05$$

or

$$z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = 0.05$$

Solve

$$1.96\sqrt{\frac{(.12)(.88)}{n}} = 0.05$$

and round up to obtain $n = 163$

Example

On October 14, 2003, the New York Times reported that a recent poll indicated that 52 percent of the population was in favor of the job performance of President Bush, with a margin of error of ± 4 percent and 95% confidence level. Can we infer **how many people were questioned**?

Solution

- $\alpha = .05, z_{.025} = 1.96$
- $\hat{p} = .52$
- $ME = z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = 1.96\sqrt{\frac{.52(.48)}{n}}$
- $1.96\sqrt{.52(.48)/n} = .04$
- $n = 599$

Practice

A sample of 100 transistors is randomly chosen from a large batch and tested to determine if they meet the current standards. If 80 of them meet the standards, then find **95% confidence interval for p, the fraction of all the transistors that meet the standards.**

Estimation of population variance σ^2 for normal population

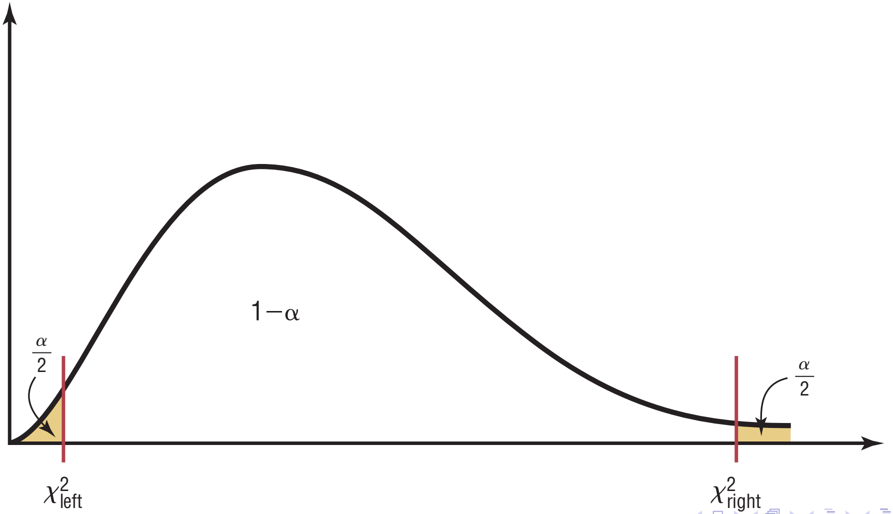
$$\frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}$$

- n: sample size
- S^2: sample variance

Estimation of population variance σ^2 for normal population

$$\frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}$$

- n: sample size
- S^2: sample variance


1 − α
α/2
α/2
χ^2_{left}
χ^2_{right}

$$P(\chi^2_{1-\alpha/2,n-1} < (n-1)\frac{S^2}{\sigma^2} < \chi^2_{\alpha/2,n-1}) = 1-\alpha$$

or

$$P(\frac{(n-1)S^2}{\chi^2_{\alpha/2,n-1}} < \sigma^2 < \frac{(n-1)S^2}{\chi^2_{1-\alpha/2,n-1}}) = 1-\alpha$$

$100(1-\alpha)\%$ CI of σ^2

$$\boxed{\frac{(n-1)S^2}{\chi^2_{\alpha/2,n-1}} < \sigma^2 < \frac{(n-1)S^2}{\chi^2_{1-\alpha/2,n-1}}}$$

Example

The sugar content of the syrup in canned peaches is normally distributed. A random sample of $n = 10$ cans yields a sample standard deviation of $s = 4.8$ milligrams. Calculate **a 95% two-sided CI for the population variance** σ^2**.**

Solution

- Information
  - sample size $n = 10$
  - sample standard deviation $s = 4.8mg$
  - Confidence level $1 - \alpha = .95 \Rightarrow \alpha/2 = .025$

- Critical value of χ^2

$$\chi^2_{1-\alpha/2,n-1} = \chi^2_{.975,9} = 2.7$$

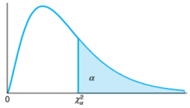


Table A.5 Critical Values of the Chi-Squared Distribution

	α									
v	**0.995**	**0.99**	**0.98**	**0.975**	**0.95**	**0.90**	**0.80**	**0.75**	**0.70**	**0.50**
1	0.0^4393	0.0^3157	0.0^3628	0.0^3982	0.00393	0.0158	0.0642	0.102	0.148	0.455
2	0.0100	0.0201	0.0404	0.0506	0.103	0.211	0.446	0.575	0.713	1.386
3	0.0717	0.115	0.185	0.216	0.352	0.584	1.005	1.213	1.424	2.366
4	0.207	0.297	0.429	0.484	0.711	1.064	1.649	1.923	2.195	3.357
5	0.412	0.554	0.752	0.831	1.145	1.610	2.343	2.675	3.000	4.351
6	0.676	0.872	1.134	1.237	1.635	2.204	3.070	3.455	3.828	5.348
7	0.989	1.239	1.564	1.690	2.167	2.833	3.822	4.255	4.671	6.346
8	1.344	1.647	2.032	2.180	2.733	3.490	4.594	5.071	5.527	7.344
9	1.735	2.088	2.532	2.700	3.325	4.168	5.380	5.899	6.393	8.343
10	2.156	2.558	3.059	3.247	3.940	4.865	6.179	6.737	7.267	9.342
11	2.603	3.053	3.609	3.816	4.575	5.578	6.989	7.584	8.148	10.341

- Critical value of χ^2

$$\chi^2_{\alpha/2,n-1} = \chi^2_{.025,9} = 19.023$$

Table A.5 (continued) Critical Values of the Chi-Squared Distribution

	α									
v	**0.30**	**0.25**	**0.20**	**0.10**	**0.05**	**0.025**	**0.02**	**0.01**	**0.005**	**0.001**
1	1.074	1.323	1.642	2.706	3.841	5.024	5.412	6.635	7.879	10.827
2	2.408	2.773	3.219	4.605	5.991	7.378	7.824	9.210	10.597	13.815
3	3.665	4.108	4.642	6.251	7.815	9.348	9.837	11.345	12.838	16.266
4	4.878	5.385	5.989	7.779	9.488	11.143	11.668	13.277	14.860	18.466
5	6.064	6.626	7.289	9.236	11.070	12.832	13.388	15.086	16.750	20.515
6	7.231	7.841	8.558	10.645	12.592	14.449	15.033	16.812	18.548	22.457
7	8.383	9.037	9.803	12.017	14.067	16.013	16.622	18.475	20.278	24.321
8	9.524	10.219	11.030	13.362	15.507	17.535	18.168	20.090	21.955	26.124
9	10.656	11.389	12.242	14.684	16.919	19.023	19.679	21.666	23.589	27.877
10	11.781	12.549	13.442	15.987	18.307	20.483	21.161	23.209	25.188	29.588

- Upper bound

$$\frac{(n-1)s^2}{\chi^2_{1-\alpha/2,n-1}} = \frac{9 * (4.8)^2}{2.7} = 76.8$$

- Lower bound

$$\frac{(n-1)s^2}{\chi^2_{\alpha/2,n-1}} = \frac{9 * (4.8)^2}{19.023} = 10.9$$

- 95% CI for population variance

$$10.9 < \sigma^2 < 76.8(mg^2)$$

The following are the weights, in decagrams, of 10 packages of grass seed distributed by a certain company: 46.4, 46.1, 45.8, 47.0, 46.1, 45.9, 45.8, 46.9, 45.2, and 46.0. Find a **95% confidence interval for the variance of the weights of all such packages of grass seed distributed by this company**, assuming a normal population

Keywords

- point estimate and efficient estimator for population mean, proportion and variance are sample mean, sample proportion and sample variance

- two-sided $100(1-\alpha)\%$ CI for population mean μ

$$(\bar{x} - ME, \qquad \bar{x} + ME)$$

  - Case 1: population variance σ^2 known, large sample size or normal population
    $ME = z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$
  - Case 2: population variance σ^2 unknown, normal population $ME = t_{\alpha/2,n-1}\frac{s}{\sqrt{n}}$

- two-sided $100(1-\alpha)\%$ CI for population proportion p

$$(\hat{p} - ME, \qquad \hat{p} + ME)$$

where

$$ME = z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

- two-sided $100(1-\alpha)\%$ CI for population variance σ^2 - normal population

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