Introduction to Estimation

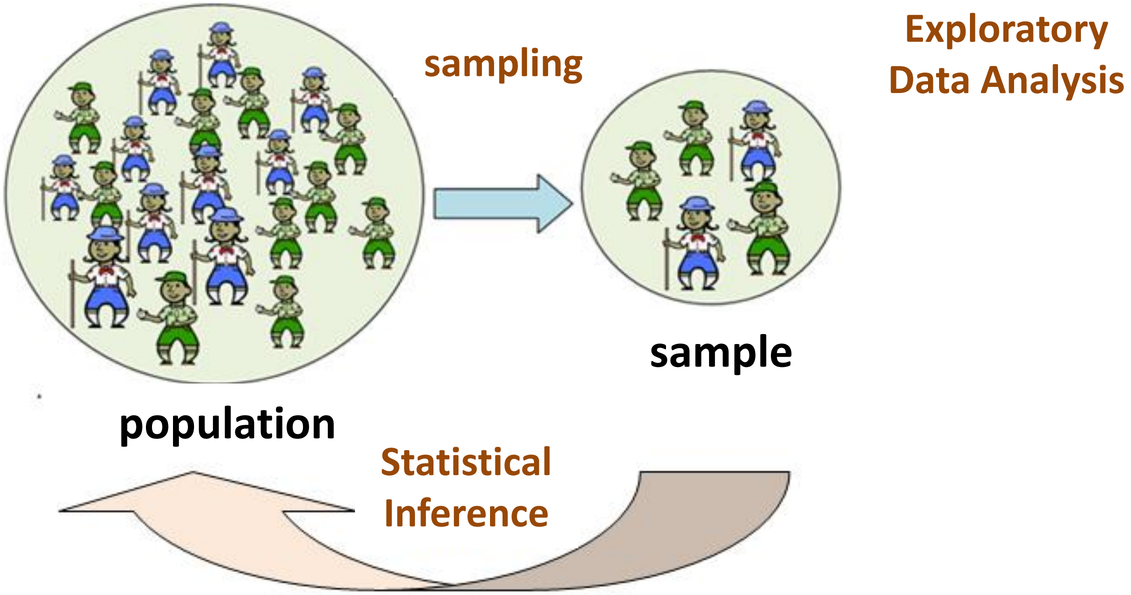
Martina Litschmannová

martina.litschmannova@vsb.cz

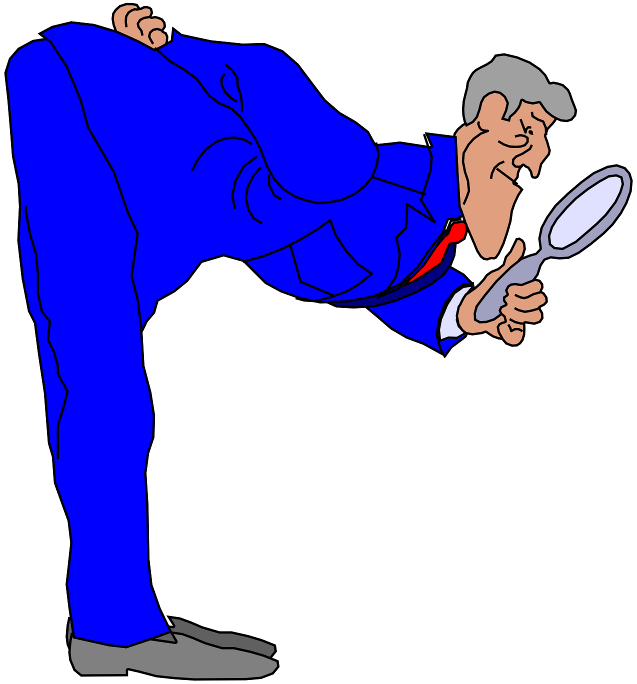
K210

Populations vs. Sample

- A **population** includes each element from the set of observations that can be made.
- A **sample** consists only of observations drawn from the population.



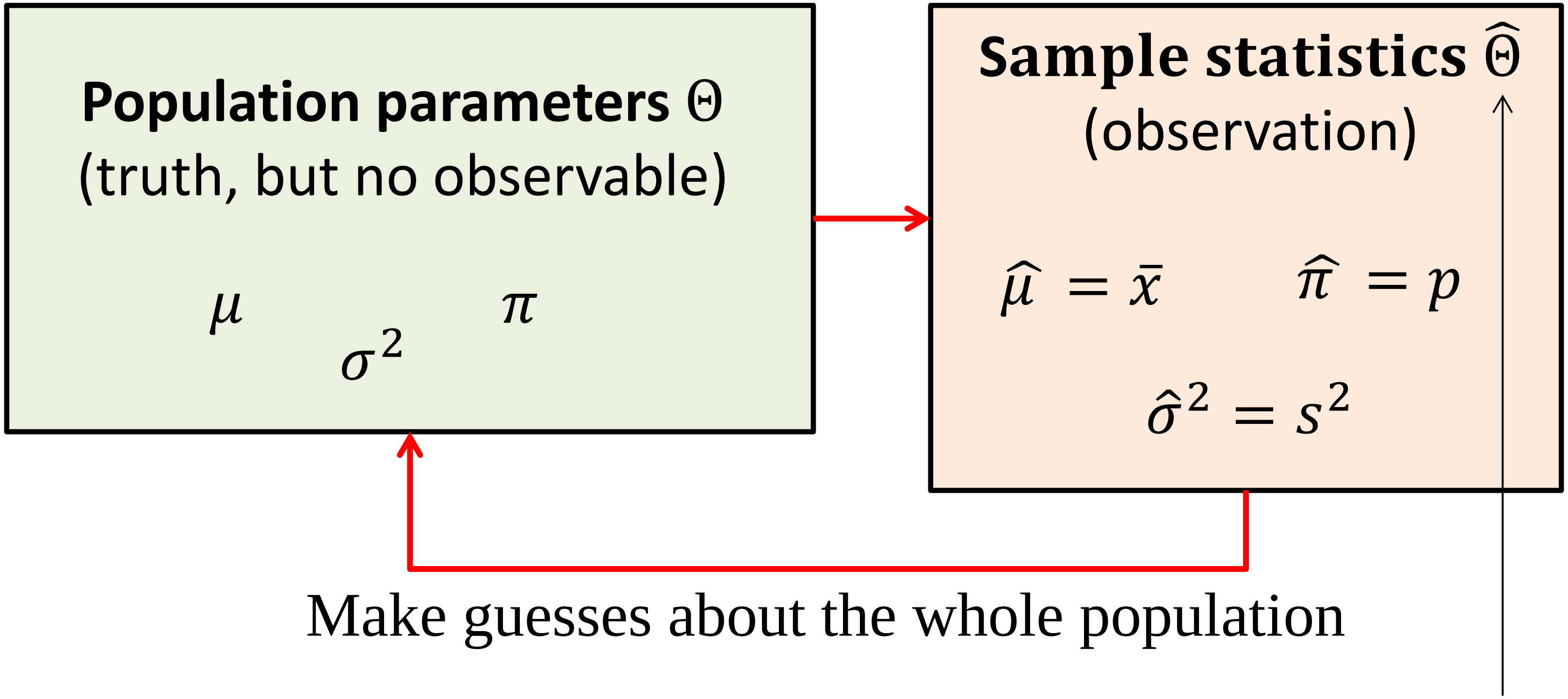
What is statistical inference?



Use a random sample to learn something about a larger population.

What is statistical inference?

The process of making guesses about the truth from a sample.



Population parameters Θ

(truth, but no observable)

μ σ^2 π

Sample statistics $\widehat{\Theta}$

(observation)

$\widehat{\mu} = \bar{x}$ $\widehat{\pi} = p$

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

Make guesses about the whole population

hat notation ^ is often used to indicate “estitmate”

Characteristic of a population
vs.
characteristic of a sample

- A a measurable characteristic of a population, such as a mean or standard deviation, is called a parameter, but a measurable characteristic of a sample is called a statistic.

Population	Expectation (mean) $E(X)$, resp. μ	Median $x_{0,5}$	Variance (dispersion) $D(X)$, resp. σ^2	Std. deviation σ	Probability π
Sample	Sample mean (average) $\bar{X}$	Sample median $\tilde{X}_{0,5}$	Sample variance S^2	Sample std. deviation S	Relative frequency p

Estimation

- There are two types of inference: estimation and hypothesis testing; estimation is introduced first.

- The objective of estimation is to determine the approximate value of a population parameter on the basis of a sample statistic.

- E.g., the sample mean ($\bar{x}$) is employed to estimate the population mean (μ).

Estimation

	Statistic		**Parameter**
Mean:	$\bar{x}$	estimates	μ
Standard deviation:	s	estimates	σ
Probability:	p	estimates	π

from sample

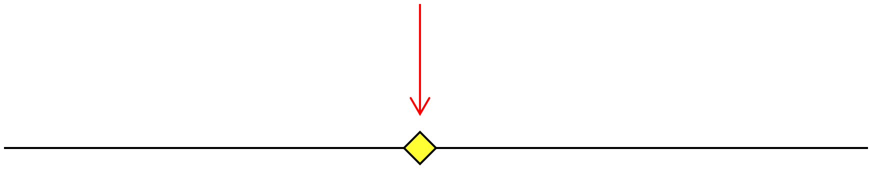
from entire population

Estimation

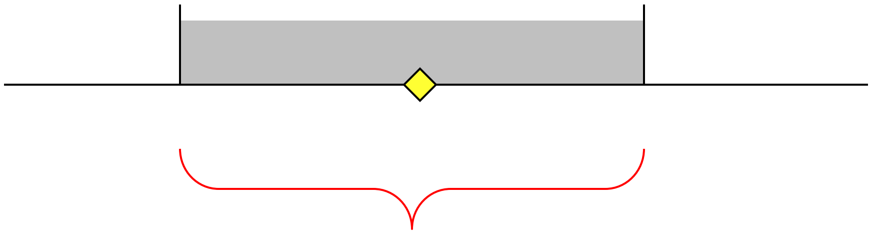
The objective of estimation is to determine the approximate value of a population parameter on the basis of a sample statistic.

There are two types of estimators:

- Point Estimator

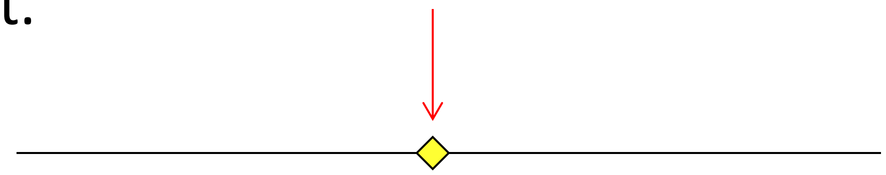


- Interval Estimator



Point Estimator

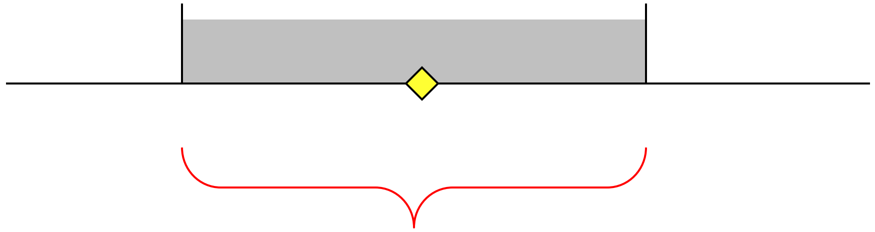
- A point estimator draws inferences about a population by estimating the value of an unknown parameter using a single value or point.



- We saw earlier that point probabilities in continuous distributions were virtually zero. Likewise, we'd expect that the point estimator gets closer to the parameter value with an increased sample size, but point estimators don't reflect the effects of larger sample sizes. Hence we will employ the interval estimator to estimate population parameters.

Interval Estimator

- An interval estimator draws inferences about a population by estimating the value of an unknown parameter using an interval.



- That is we say (with some ?? % certainty) that the population parameter of interest is between some lower and upper bounds.

Point & Interval Estimation

For example, suppose we want to estimate the mean summer income of VSB-TUO students. For n=25 students, is $\bar{x}$ calculated to be 400 $/week.

interval estimation

point estimation

- An alternative statement is:

The mean income is **between** 380 and 420 $/week.

Qualities of Estimators

Statisticians have already determined the “best” way to estimate a population parameter. Qualities desirable in estimators include unbiasedness, consistency, and relative efficiency:

- An unbiased estimator of a population parameter is an estimator whose expected value is equal to that parameter.
- An unbiased estimator is said to be consistent if the difference between the estimator and the parameter grows smaller as the sample size grows larger.
- If there are two unbiased estimators of a parameter, the one whose variance is smaller is said to be relatively efficient.

Confidence Interval Estimator for μ

Assumption: sampling distribution of the statistic is normal or nearly normal.

The central limit theorem states that the sampling distribution of a statistic will be normal or nearly normal, if any of the following conditions apply.

- The population distribution is normal.
- The sampling distribution is symmetric, unimodal, without outliers, and the sample size is 15 or less.
- The sampling distribution is moderately skewed, unimodal, without outliers, and the sample size is between 16 and 40.
- The sample size is greater than 40, without outliers.

Confidence Interval Estimator for μ

$$P\left(\bar{x} - t_{1-\alpha/_{2};n-1}\frac{s}{\sqrt{n}} < \mu < \bar{x} + t_{1-\alpha/_{2};n-1}\frac{s}{\sqrt{n}}\right) = 1 - \alpha$$

Confidence Interval Estimator for μ

The probability $1 - \alpha$ is called the confidence level.

$$\bar{x} \pm t_{1-\alpha/2;n-1} \frac{s}{\sqrt{n}}$$

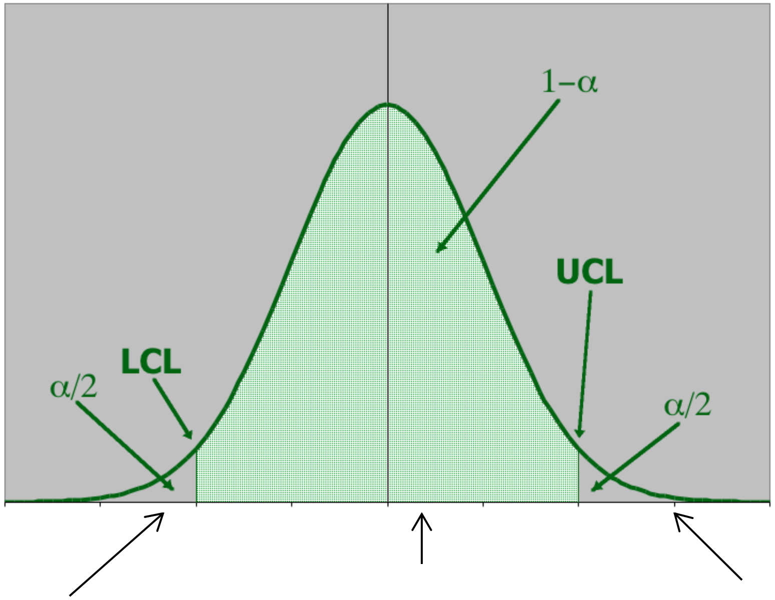
Confidence Interval Estimator for μ

The probability $1-\alpha$ is called the confidence level.

$$\bar{x} \pm t_{1-{\alpha}/{2};n-1}\frac{s}{\sqrt{n}} = \left\{\underbrace{\bar{x} - t_{1-{\alpha}/{2};n-1}\frac{s}{\sqrt{n}}}_{\text{Lower Confidence Limit - LCL}};\underbrace{\bar{x} + t_{1-{\alpha}/{2};n-1}\frac{s}{\sqrt{n}}}_{\text{Upper Confidence Limit - UCL}}\right\}$$

Graphically

- The actual location of the population mean…



…may be here… …or here… …or possibly even here…

The population mean is a fixed but unknown quantity. Its incorrect to interpret the confidence interval estimate as a probability statement about μ. The interval acts as the lower and upper limits of the interval estimate of the population mean.

1. A computer company samples demand during lead time over 25 time periods:

235	374	309	499	253
421	361	514	462	369
394	439	348	344	330
261	374	302	466	535
386	316	296	332	334

We want to estimate the mean demand over lead time with 95% confidence in order to set inventory levels.

- We want to estimate the mean demand over lead time with 95% confidence in order to set inventory levels.

IDENTIFY

- The parameter to be estimated is the pop'n mean μ.

- Confidence interval estimator will be: $\bar{x} \pm t_{1-\alpha/_{2};n-1} \frac{s}{\sqrt{n}}$

CALCULATE

- In order to use our confidence interval estimator, we need the following pieces of data:

$\bar{x}$	370,2
$t_{1-{\alpha}/{2};n-1}$	2,1
s	80,8
n	25

Calculated from the data

- therefore: $\bar{x} \pm t_{1-{\alpha}/{2};n-1} \frac{s}{\sqrt{n}} = 370{,}2 \pm 2{,}1 \cdot \frac{80{,}8}{\sqrt{25}} = 370{,}2 \pm 33{,}3$
- The lower and upper confidence limits are **336,7** and **399,5**.

CALCULATE

- In order to use our confidence interval estimator, we need the following pieces of data:

$\bar{x}$	370,2
$t_{1-\alpha/_2;n-1}$	2,1
s	80,8
n	25

Calculated from the data

CONFIDENCE.T(α;s;n)

- therefore: $\bar{x} \pm t_{1-\alpha/_2;n-1}\frac{s}{\sqrt{n}} = 370{,}2 \pm 2{,}1 \cdot \frac{80{,}8}{\sqrt{25}} = 370{,}2 \pm 33{,}3$
- The **lower** and **upper** confidence limits are **336,7** and **399,5**.

CALCULATE

- In order to use our confidence interval estimator, we need the following pieces of data:

$\bar{x}$	370,2
$t_{1-\alpha/_{2};n-1}$	2,1
s	80,8
n	25

Calculated from the data

CONFIDENCE.T(α;s;n)

- therefore: $\bar{x} \pm t_{1-\alpha/_{2};n-1}\frac{s}{\sqrt{n}} = 370{,}2 \pm 2{,}1 \cdot \frac{80{,}8}{\sqrt{25}} = 370{,}2 \pm 33{,}3$

$$P(336{,}7 < \mu < 399{,}5) = 0{,}95$$

Interval Width

The width of the confidence interval estimate is a function of the confidence level, the sample standard deviation, and the sample size.

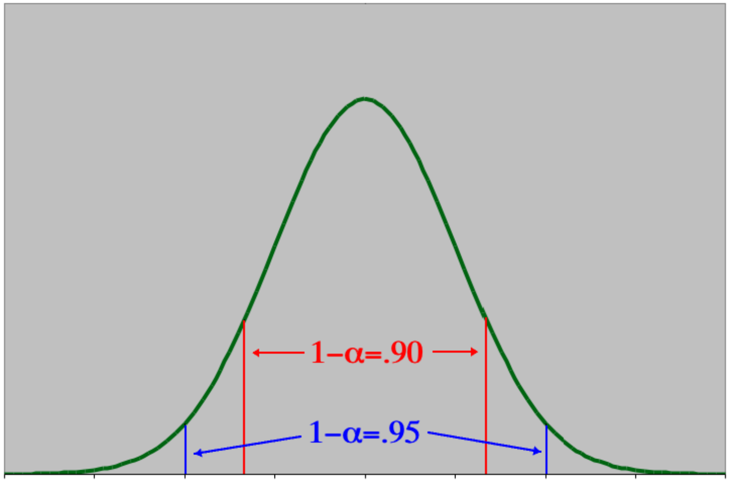
$$\bar{x} \pm t_{1-\alpha/_{2};n-1} \frac{s}{\sqrt{n}}$$

Interval Width

The width of the confidence interval estimate is a function of the confidence level, the sample standard deviation, and the sample size.

$$\bar{x} \pm t_{1-\alpha/_{2};n-1} \frac{s}{\sqrt{n}}$$

A larger confidence level produces a **wider** confidence interval.

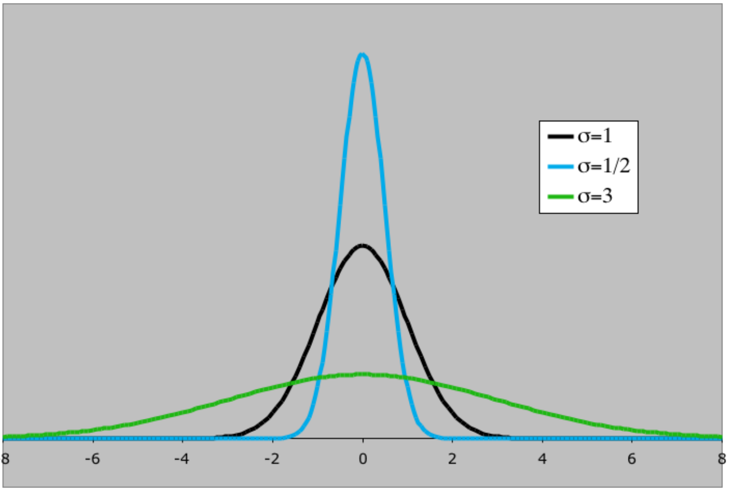


Interval Width

The width of the confidence interval estimate is a function of the confidence level, the sample standard deviation, and the sample size.

$$\bar{x} \pm t_{1-\alpha/_{2};n-1} \frac{s}{\sqrt{n}}$$

A larger standard deviation produces a **wider** confidence interval.



Interval Width

The width of the confidence interval estimate is a function of the confidence level, the sample standard deviation, and the sample size.

$$\bar{x} \pm t_{1-\alpha/_{2};n-1} \frac{s}{\sqrt{n}}$$

- Increasing the sample size **decreases** the width of the confidence interval while the confidence level can remain unchanged.

Sample Size to Estimate a Mean

- The general formula for the sample size needed to estimate a population mean with an interval estimate of:

$$\bar{x} \pm t_{1-{\alpha}/{2};n-1} \frac{s}{\sqrt{n}} = \bar{x} \pm ME$$

- Requires a sample size of at least this large:

$$n \geq \left(t_{1-{\alpha}/{2};n-1} \frac{s}{ME_{max}} \right)^2$$

2. A lumber company must estimate the mean diameter of trees to determine whether or not there is sufficient lumber to harvest an area of forest. They need to estimate this to within 1 inch at a confidence level of 99%. The tree diameters are normally distributed with a standard deviation of 6 inches.

How many trees need to be sampled?

Estimation problems

Statistic	Assumptions	Critical value	Standard Error
Sample mean, $\bar{x}$	normality, large sample	$z_{1-\alpha/2}$	$\frac{s}{\sqrt{n}}$
	normality	$t_{1-\alpha/2;n-1}$	$\frac{s}{\sqrt{n}}$
Sample proportion, p	$n > \frac{9}{p(1-p)}$	$z_{1-\alpha/2}$	$\sqrt{\frac{p(1-p)}{n}}$
Difference between means, $\bar{x}_1 - \bar{x}_2$	normality, large samples	$z_{1-\alpha/2}$	$\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$
	normality	$t_{1-\alpha/2;DF}$ $DF = \frac{\left(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}\right)^2}{\frac{\left(\frac{S_1^2}{n_1}\right)^2}{n_1 - 1} + \frac{\left(\frac{S_2^2}{n_2}\right)^2}{n_2 - 1}}$	$\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$
Difference between proportions, $p_1 - p_2$	$\forall i \in \{1,2\}$: $n_i > 30$, $n_i > \frac{9}{p_i(1-p_i)}$	$z_{1-\alpha/2}$	$\sqrt{\frac{p_1(1-p_1)}{n} + \frac{p_2(1-p_2)}{n}}$

3. Suppose a simple random sample of 150 students is drawn from a population of 3000 college students. Among sampled students, the average IQ score is 115 with a standard deviation of 10. What is the 99% confidence interval for the students' IQ score?

(A) 115 ± 0.01
(B) 115 ± 0.82
(C) 115 ± 2.1
(D) 115 ± 2.6
(E) None of the above

4. Suppose that simple random samples of college freshman are selected from two universities - 15 students from school A and 20 students from school B. On a standardized test, the sample from school A has an average score of 1000 with a standard deviation of 100. The sample from school B has an average score of 950 with a standard deviation of 90. What is the 90% confidence interval for the difference in test scores at the two schools, assuming that test scores came from normal distributions in both schools?

(A) 50 ± 1.70
(B) 50 ± 28.49
(C) 50 ± 32.74
(D) 50 ± 55.66
(E) None of the above

Estimate the mean difference between matched data pairs

5. Twenty-two students were randomly selected from a population of 1000 students. The sampling method was simple random sampling. All of the students were given a standardized English test and a standardized math test. Test results are in dataset test.xls.

Find the 90% confidence interval for the mean difference between student scores on the math and English tests. Assume that the mean differences are approximately normally distributed.

See at http://stattrek.com/estimation/mean-difference-pairs.aspx?Tutorial=AP.

6. A major metropolitan newspaper selected a simple random sample of 1,600 readers from their list of 100,000 subscribers. They asked whether the paper should increase its coverage of local news. Forty percent of the sample wanted more local news. What is the 99% confidence interval for the proportion of readers who would like more coverage of local news?

(A) 0.30 to 0.50
(B) 0.32 to 0.48
(C) 0.35 to 0.45
(D) 0.37 to 0.43
(E) 0.39 to 0.41

7. Suppose the Cartoon Network conducts a nation-wide survey to assess viewer attitudes toward Superman. Using a simple random sample, they select 400 boys and 300 girls to participate in the study. Forty percent of the boys say that Superman is their favorite character, compared to thirty percent of the girls. What is the 90% confidence interval for the true difference in attitudes toward Superman?

(A) 0 to 20 percent more boys prefer Superman
(B) 2 to 18 percent more boys prefer Superman
(C) 4 to 16 percent more boys prefer Superman
(D) 6 to 14 percent more boys prefer Superman
(E) None of the above

Study materials :

- http://homel.vsb.cz/~bri10/Teaching/Bris%20Prob%20&%20Stat.pdf
  (p. 130 - p.141)

- http://stattrek.com/tutorials/ap-statistics-tutorial.aspx
  (Statistical Inference –Estimation, Estimation Problem)