

Chapter 6. Point Estimation

Chapter 7. Interval Estimation

Math 3215 Summer 2023

Georgia Institute of Technology

Section 6.3.

Order Statistics

X_1, X_2, \dots, X_n i.i.d. —

$u(X_1, X_2, \dots, X_n)$: a statistic

Example : $\bar{X} = \frac{1}{n} (X_1 + \dots + X_n)$, $\frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n}$,

$\min \{ X_1, \dots, X_n \}$

$\max \{ X_1, \dots, X_n \}$

median $\{ \text{---} \}$

quantiles

percentiles

Order Statistics

Order statistics are the observations of the random sample, arranged, or ordered, in magnitude from the smallest to the largest.

Example

The values $x_1 = 0.62$, $x_2 = 0.98$, $x_3 = 0.31$, $x_4 = 0.81$, and $x_5 = 0.53$ are the $n = 5$ observed values of five independent trials of an experiment with pdf $f(x) = 2x$ for $0 < x < 1$.

The observed order statistics are $y_1 = 0.31$, $y_2 = 0.53$, $y_3 = 0.62$, $y_4 = 0.81$, $y_5 = 0.98$

The sample median is $y_3 = 0.62$

The sample range is $y_5 - y_1 = 0.98 - 0.31$.
 $[y_1, y_5]$

Order Statistics

$$F(x) = \begin{cases} 0 & , x < 0 \\ x^2 & , 0 \leq x < 1 \\ 1 & , x \geq 1 \end{cases}$$



Let X_1, X_2, X_3, X_4, X_5 be independent random variables with density

$f(x) = 2x$ for $0 < x < 1$.

$$P(X_1 = X_2) = 0 = \iint_A f(x_1) f(x_2) dx_1 dx_2$$

Let $Y_1 \leq Y_2 \leq Y_3 \leq Y_4 \leq Y_5$ be the order statistics of $X_i, i = 1, 2, 3, 4, 5$.

What is the distribution of Y_1 ?

$$Y_1 = \min \{ X_i : i = 1, 2, \dots, 5 \}$$

$$F_{Y_1}(y) = P(Y_1 \leq y) = P(\min \{ X_i : i = 1, \dots, 5 \} \leq y)$$

$$= 1 - P(\min \{ X_i \} > y)$$

$$= 1 - P(X_1 > y \text{ and } X_2 > y \text{ and } \dots)$$

$$= 1 - \underbrace{P(X_1 > y)}_{(1-F(y))} \underbrace{P(X_2 > y)}_{(1-F(y))} \dots \underbrace{P(X_5 > y)}_{(1-F(y))}$$

$$= 1 - (1 - F(y))^5$$

$$f_{Y_1}(y) = -5(1 - F(y))^4 \cdot (-f(y))$$

$$= 5 \cdot (1 - F(y))^4 \cdot f(y)$$

Order Statistics

Let X_1, X_2, X_3, X_4, X_5 be independent random variables with density $f_X(x) = 2x$ for $0 < x < 1$.

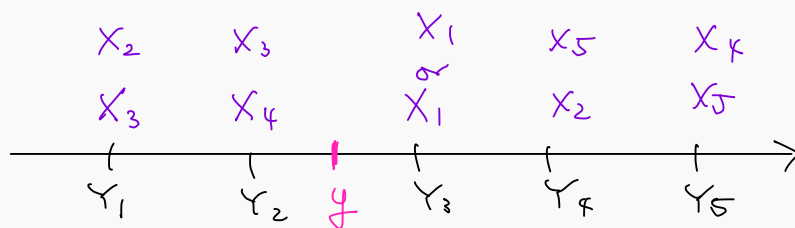
Let $Y_1 \leq Y_2 \leq Y_3 \leq Y_4 \leq Y_5$ be the order statistics of $X_i, i = 1, 2, 3, 4, 5$.

What is the distribution of Y_5 ?

$$\begin{aligned} F_{Y_5}(y) &= P(Y_5 \leq y) = P(\max\{X_i\} \leq y) \\ &= P(X_1 \leq y \text{ and } X_2 \leq y \text{ and } \dots X_5 \leq y) \\ &= P(X_1 \leq y) P(X_2 \leq y) \dots P(X_5 \leq y) \\ &= \underbrace{F(y) \cdot F(y) \cdot \dots F(y)}_{5 \text{ times}} = F(y)^5 \end{aligned}$$

$$f_{Y_5}(y) = 5 \cdot F_X(y)^4 \cdot f_X(y).$$

Order Statistics



Let X_1, X_2, X_3, X_4, X_5 be independent random variables with density $f(x) = 2x$ for $0 < x < 1$.

Let $Y_1 \leq Y_2 \leq Y_3 \leq Y_4 \leq Y_5$ be the order statistics of $X_i, i = 1, 2, 3, 4, 5$.

What is the distribution of Y_2 ?

$$\begin{aligned}
 F_{Y_2}(y) &= P(Y_2 \leq y) = P(Y_1 \leq y, Y_2 \leq y, Y_3 > y, Y_4 > y, Y_5 > y) \\
 &= P(X_3 \leq y, X_4 \leq y, X_1 > y, X_2 > y, X_5 > y) \\
 &\quad \cup \\
 &\quad \cup \\
 &\quad \cup \\
 &\quad \cup \\
 &\quad \cup \\
 &= P(\text{---}) + P(\text{---}) + \dots + \dots \\
 &= P(X_3 \leq y) \cdot P(X_4 \leq y) \cdots P(X_5 > y) + \dots \\
 &= \binom{5}{2} F(y) \cdot F(y) (1 - F(y))(1 - F(y))(1 - F(y)) = \binom{5}{2} F(y)^2 (1 - F(y))^3
 \end{aligned}$$

$$F_{Y_2}(y) = P(Y_2 \leq y)$$

$$= P(Y_2 \leq y, Y_3 > y) + P(Y_3 \leq y, Y_4 > y)$$

$$+ P(Y_4 \leq y, Y_5 > y)$$

$$+ P(Y_5 \leq y)$$

$$= \binom{5}{2} F(y)^2 (1-F(y))^3 + \binom{5}{3} F(y)^3 (1-F(y))^2$$

$$+ \binom{5}{4} F(y)^4 (1-F(y)) + \binom{5}{5} F(y)^5$$

$$= \sum_{k=2}^5 \binom{5}{k} F(y)^k (1-F(y))^{5-k}$$

Order Statistics

In general, if $Y_1 < Y_2 < \dots < Y_n$ are the order statistics of independent random variables X_i for $i = 1, 2, \dots, n$ with pdf f ,

then the distribution of Y_i is

$$F_{Y_i}(y) = \sum_{k=i}^n \binom{n}{k} \underbrace{F(y)^k}_{k-1} \underbrace{(1-F(y))^{n-k}}_{n-k}$$

$$\begin{aligned} f_{Y_i}(y) &= \frac{d}{dy} F_{Y_i}(y) \\ &= \sum_{k=i}^n \binom{n}{k} \frac{d}{dy} \left(F(y)^k (1-F(y))^{n-k} \right) \\ &= \sum_{k=i}^n \binom{n}{k} \left(k F(y)^{k-1} (1-F(y))^{n-k} - F(y)^k (n-k) (1-F(y))^{n-k-1} \right) \\ &= \sum_{k=i}^n \binom{n}{k} F(y)^{k-1} (1-F(y))^{n-k} \left(k - (n-k) \frac{F(y)}{1-F(y)} \right) \\ &= \frac{n!}{(i-1)! (n-i)!} (F(y))^{i-1} (1-F(y))^{n-i} f(y) \end{aligned}$$

Order Statistics

Example

Let X_1, X_2, X_3, X_4, X_5 be independent random variables with density $f(x) = 2x$ for $0 < x < 1$.

Let $Y_1 \leq Y_2 \leq Y_3 \leq Y_4 \leq Y_5$ be the order statistics of $X_i, i = 1, 2, 3, 4, 5$.

Find $\mathbb{P}(Y_3 \leq y)$.

Find the density of Y_3 .

$$\mathbb{P}(Y_3 \leq y) = \sum_{k=3}^5 \binom{5}{k} (y^2)^k \cdot (1-y^2)^{5-k}, \quad 0 \leq y \leq 1$$

$$= 0, \quad y < 0$$

$$= 1, \quad y \geq 1$$

$$f_{Y_3}(y) = \frac{5!}{(3-1)!(5-3)!} (y^2)^{3-1} (1-y^2)^{5-3} \cdot (2y)$$

$$= 30 \cdot y^4 (1-y^2)^2 \cdot 2y \quad 0 \leq y \leq 1$$

Exercise

Let $Y_1 < Y_2 < \dots < Y_{19}$ be the order statistics of $n = 19$ independent observations from the exponential distribution with mean θ .

Find the pdf of Y_1 .

Find $\mathbb{E}[F(Y_1)]$ where F is the cdf of the exponential distribution.

5.6-7. Let X equal the maximal oxygen intake of a human on a treadmill, where the measurements are in milliliters of oxygen per minute per kilogram of weight. Assume that, for a particular population, the mean of X is $\mu = 54.030$ and the standard deviation is $\sigma = 5.8$. Let \bar{X} be the sample mean of a random sample of size $n = 47$. Find $P(52.761 \leq \bar{X} \leq 54.453)$, approximately.

5.6-8. Let X equal the weight in grams of a miniature candy bar. Assume that $\mu = E(X) = 24.43$ and $\sigma^2 = \text{Var}(X) = 2.20$. Let \bar{X} be the sample mean of a random sample of $n = 30$ candy bars. Find

(a) $E(\bar{X})$. (b) $\text{Var}(\bar{X})$. (c) $P(24.17 \leq \bar{X} \leq 24.82)$, approximately.

5.6-9. In Example 5.6-4, with $n = 4$, compute $P(1.7 \leq Y \leq 3.2)$ and compare your answer with the normal approximation of this probability.

5.6-10. Let X and Y equal the respective numbers of hours a randomly selected child watches movies or cartoons on TV during a certain month. From experience, it is known that $E(X) = 30$, $E(Y) = 50$, $\text{Var}(X) = 52$, $\text{Var}(Y) = 64$, and $\text{Cov}(X, Y) = 14$. Twenty-five children are selected at random. Let Z equal the total number of hours these 25 children watch TV movies or cartoons in the next month. Approximate $P(1970 < Z < 2090)$. **HINT:** Use the remark after Theorem 5.3-2.

5.6-11. The tensile strength X of paper, in pounds per square inch, has $\mu = 30$ and $\sigma = 3$. A random sample of size $n = 100$ is taken from the distribution of tensile

strengths. Compute the probability that the sample mean \bar{X} is greater than 29.5 pounds per square inch.

5.6-12. At certain times during the year, a bus company runs a special van holding ten passengers from Iowa City to Chicago. After the opening of sales of the tickets, the time (in minutes) between sales of tickets for the trip has a gamma distribution with $\alpha = 3$ and $\theta = 2$.

(a) Assuming independence, record an integral that gives the probability of being sold out within one hour.

(b) Approximate the answer in part (a).

5.6-13. Let X_1, X_2, X_3, X_4 represent the random times in days needed to complete four steps of a project. These times are independent and have gamma distributions with common $\theta = 2$ and $\alpha_1 = 3, \alpha_2 = 2, \alpha_3 = 5, \alpha_4 = 3$, respectively. One step must be completed before the next can be started. Let Y equal the total time needed to complete the project.

(a) Find an integral that represents $P(Y \leq 25)$.

(b) Using a normal distribution, approximate the answer to part (a). Is this approach justified?

5.6-14. Suppose that the sick leave taken by the typical worker per year has $\mu = 10$, $\sigma = 2$, measured in days. A firm has $n = 20$ employees. Assuming independence, how many sick days should the firm budget if the financial officer wants the probability of exceeding the number of days budgeted to be less than 20%?

5.7 APPROXIMATIONS FOR DISCRETE DISTRIBUTIONS

In this section, we illustrate how the normal distribution can be used to approximate probabilities for certain discrete-type distributions. One of the more important discrete distributions is the binomial distribution. To see how the central limit theorem can be applied, recall that a binomial random variable can be described as the sum of Bernoulli random variables. That is, let X_1, X_2, \dots, X_n be a random sample from a Bernoulli distribution with mean $\mu = p$ and variance $\sigma^2 = p(1-p)$, where $0 < p < 1$. Then $Y = \sum_{i=1}^n X_i$ is $b(n, p)$. The central limit theorem states that the distribution of

$$W = \frac{Y - np}{\sqrt{np(1-p)}} = \frac{\bar{X} - p}{\sqrt{p(1-p)/n}}$$

is $N(0, 1)$ in the limit as $n \rightarrow \infty$. Thus, if n is “sufficiently large,” the distribution of Y is approximately $N[np, np(1-p)]$, and probabilities for the binomial distribution $b(n, p)$ can be approximated with this normal distribution. A rule often stated is that n is sufficiently large if $np \geq 5$ and $n(1-p) \geq 5$.

Note that we shall be approximating probabilities for a discrete distribution with probabilities for a continuous distribution. Let us discuss a reasonable procedure in this situation. If V is $N(\mu, \sigma^2)$, then $P(a < V < b)$ is equivalent to the area bounded by the pdf of V , the v -axis, $v = a$, and $v = b$. Now recall that for a Y that is $b(n, p)$,

$$W = X + Y$$

$$\bar{W} = \frac{1}{n} (\overset{W_1}{\underbrace{(X_1 + Y_1)}} + \dots + \overset{W_n}{\underbrace{(X_n + Y_n)}})$$

$$= \bar{X} + \bar{Y}$$

$$E[\bar{W}] = E[\bar{X}] + E[\bar{Y}] = EX + EY = 80$$

$$\text{Var}(\bar{W}) = \text{Var}(\bar{X} + \bar{Y}) = \frac{144}{25} = \frac{\sigma^2}{n}$$

$$\begin{aligned} \text{Var}(W_1) &= \text{Var}(X + Y) = \underbrace{\text{Var}(X)} + \underbrace{\text{Var}(Y)} + 2 \underbrace{\text{Cov}(X, Y)} \\ &= 52 + 64 + 2 \cdot 14 \\ &= 144 = \sigma^2 \end{aligned}$$

$$\frac{\bar{W} - 80}{\sqrt{144/25}} \Rightarrow N(0, 1)$$

CLT :

$$\frac{\bar{X} - \mu}{\underbrace{\sigma/\sqrt{n}}_{\text{"}} \sqrt{\text{Var}(\bar{X})}} = \frac{\bar{X} - E[\bar{X}]}{\sqrt{\text{Var}(\bar{X})}}$$

$$\sigma = \sqrt{\text{Var}(X)}$$

Section 6.4.

Maximum Likelihood and Method of Moments Estimation

Estimators

We consider an experiment whose outcome follows a certain distribution.

Assume that the distribution is characterized by a pdf $f(x; \theta)$ where θ is an unknown parameter.

Let Ω be the set of all possible parameters. We call it the parameter space.

Suppose we repeat the n independent experiment and observe the outcome (sample) X_1, X_2, \dots, X_n .

Our goal is to estimate θ based on the samples.

Estimators

Definition

For a function $u : \mathbb{R}^n \rightarrow \mathbb{R}$, $u(X_1, X_2, \dots, X_n)$ is called an estimator (or a point estimator) of θ .

Estimators

Example

Suppose X_1, X_2, \dots, X_n are independent Bernoulli RVs with unknown success probability $p \in [0, 1]$.

We want to find an estimator of p which maximizes its possibility.

Maximum Likelihood Estimators

Definition

Let X_1, X_2, \dots, X_n be a random sample from a distribution $f(x; \theta_1, \dots, \theta_m)$, $(\theta_1, \dots, \theta_m) \in \Omega$.

The likelihood function of $\theta_1, \dots, \theta_m$ is

The maximum likelihood estimators of $\theta_1, \dots, \theta_m$ are

Maximum Likelihood Estimators

Example

Suppose X_1, X_2, \dots, X_n are independent exponential with mean $\theta \in (0, \infty)$.

Maximum Likelihood Estimators

Example

Suppose X_1, X_2, \dots, X_n are normal with mean $\theta_1 \in \mathbb{R}$ and variance $\theta_2 \in (0, \infty)$.

Unbiased estimators

Definition

If an estimator $u(X_1, X_2, \dots, X_n)$ of θ satisfies $\mathbb{E}[u(X_1, X_2, \dots, X_n)] = \theta$, we say the estimator is unbiased.

Unbiased estimators

Example

Suppose X_1, X_2, \dots, X_n are normal with mean $\theta_1 \in \mathbb{R}$ and variance $\theta_2 \in (0, \infty)$.

Exercise

A random sample X_1, X_2, \dots, X_n of size n is taken from a Poisson distribution with a mean of $\theta \in (0, \infty)$.

Find the MLE for θ .

Section 7.1.

Confidence Intervals for Means

Two-sided confidence intervals

Consider a random sample X_1, X_2, \dots, X_n from a normal distribution $N(\mu, \sigma^2)$.

We are interested in the closeness of the maximum likelihood estimator \bar{X} to the unknown mean μ .

Suppose σ^2 is known.

Our goal is to find an interval $[a, b]$ such that $\mathbb{P}(\bar{X} \in [a, b]) = 1 - \alpha$, where $\alpha > 0$.

Let $z_{\frac{\alpha}{2}} > 0$ be such that $\mathbb{P}(Z \geq z_{\frac{\alpha}{2}}) = \frac{\alpha}{2}$. (Here, Z is standard normal.)

Two-sided confidence intervals

Since $(\bar{X} - \mu)/(\sigma/\sqrt{n})$ is standard normal, we have

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

Solving for \bar{X} , we get

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

Two-sided confidence intervals

Definition

The interval $\left[\mu - z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}, \mu + z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \right]$ is called the two-sided confidence interval for unknown μ with confidence coefficient $1 - \alpha$ (for $100(1 - \alpha)\%$).

Two-sided confidence intervals

Example

Let X equal the length of life of a 60-watt light bulb marketed by a certain manufacturer.

Assume that the distribution of X is $N(\mu, 1296)$.

If a random sample of $n = 27$ bulbs is tested until they burn out, yielding a sample mean of $\bar{X} = 1478$ hours,

then what is a 95% confidence interval for μ ?

Two-sided confidence intervals

If we cannot assume that the distribution from which the sample arose is normal, we can still obtain an approximate confidence interval for μ .

By the central limit theorem, provided that n is large enough, the ratio $(\bar{X} - \mu)/(\sigma/\sqrt{n})$ has the approximate normal distribution $N(0, 1)$ when the underlying distribution is not normal.

Two-sided confidence intervals

Example

Let X equal the amount of orange juice (in grams per day) consumed by an American. Suppose it is known that the standard deviation of X is $\sigma = 96$.

An orange growers' association took a random sample of $n = 576$ Americans and found that they consumed, on average, $\bar{X} = 133$ grams of orange juice per day.

When σ^2 is unknown

Even though σ^2 is unknown and the sample distribution is not normal, if n is large enough ($n \geq 30$), then

$$\frac{\bar{X} - \mu}{\frac{S}{\sqrt{n}}}, \quad S^2 = \frac{\sum (X - \bar{X})^2}{n - 1}$$

is approximately the standard normal.

Thus, a $100(1 - \alpha)\%$ confidence interval is

If the sample distribution is normal, then $\frac{\bar{X} - \mu}{\frac{S}{\sqrt{n}}}$ has a student's t distribution with degree of freedom $n - 1$.

When σ^2 is unknown

Example

Let X equal the amount of butterfat in pounds produced by a typical cow during a 305-day milk production period between her first and second calves.

Assume that the distribution of X is $N(\mu, \sigma^2)$.

To estimate, a farmer measured the butterfat production for $n = 20$ cows and $\bar{X} = 507.50$ and $S = 89.75$.

Exercise

A random sample of size 16 from the normal distribution $N(\mu, 25)$ yielded $\bar{X} = 73.8$.

Find a 95% confidence interval for μ .

Section 7.2.

Confidence Intervals for the Difference of Two Means

Confidence Intervals for the Difference of Two Means

Let X_1, X_2, \dots, X_{n_X} and Y_1, Y_2, \dots, Y_{n_Y} be, respectively, two independent random samples of sizes n_X and n_Y from the two normal distributions $N(\mu_X, \sigma_X^2)$ and $N(\mu_Y, \sigma_Y^2)$.

The distribution of $W = \bar{X} - \bar{Y}$ is

The confidence interval for $\mu_X - \mu_Y$ is

This also works when

Confidence Intervals for the Difference of Two Means

Example

The length of life of brand X light bulbs is assumed to be $N(\mu_X, 784)$.

The length of life of brand Y light bulbs is assumed to be $N(\mu_Y, 627)$ and independent of X .

If a random sample of $n_X = 56$ brand X light bulbs yielded a mean of $\bar{X} = 937.4$ hours and a random sample of size $n_Y = 57$ brand Y light bulbs yielded a mean of $\bar{Y} = 988.9$ hours,

find a 90% confidence interval for $\mu_X - \mu_Y$.

Confidence Intervals for the Difference of Two Means

Suppose σ_X^2 and σ_Y^2 are unknown and the sample sizes are small.

Assume that $\sigma_X^2 = \sigma_Y^2$.

Then,

$$T = \frac{Z}{\sqrt{U/(n_X + n_Y - 2)}}$$

has a student's t distribution of degree of freedom $n_X + n_Y - 2$ where

$$Z = \frac{\bar{X} - \bar{Y} - (\mu_X - \mu_Y)}{\sqrt{\sigma^2/n_X + \sigma^2/n_Y}}, \quad U = \frac{(n_X - 1)S_X^2}{\sigma^2} + \frac{(n_Y - 1)S_Y^2}{\sigma^2}.$$

After simplification, T can be written as

Confidence Intervals for the Difference of Two Means

Example

Suppose that scores on a standardized test in mathematics taken by students from large and small high schools are $N(\mu_X, \sigma^2)$ and $N(\mu_Y, \sigma^2)$, respectively, where σ unknown.

If a random sample of $n_X = 9$ students from large high schools yielded $\bar{X} = 81.31$ and $S_X^2 = 60.76$, and a random sample of $n_Y = 15$ students from small high schools yielded $\bar{Y} = 78.61$ and $S_Y^2 = 48.24$.

Find a 95% confidence interval for $\mu_X - \mu_Y$.

Section 7.3.

Confidence Intervals for Proportions

Confidence Intervals for Proportions

Example

In a certain political campaign, one candidate has a poll taken at random among the voting population.

The results are that $y = 185$ out of $n = 351$ voters favor this candidate.

Should the candidate feel very confident of winning?

Confidence Intervals for Proportions

In general, suppose Y is a random sample from a binomial distribution with the given number of trials n and an unknown success probability p .

Note that Y/n is an unbiased point estimator for p and

$$\frac{(Y/n) - p}{\sqrt{p(1-p)/n}}$$

has an approximate normal distribution $N(0, 1)$.

One-sided confidence interval

One-sided confidence intervals are sometimes appropriate for p .

For example, we may be interested in an upper bound on the proportion of defectives in manufacturing some item.

For an upper bound, we consider

$$\mathbb{P} \left(-z_{\alpha} \leq \frac{(Y/n) - p}{\sqrt{(Y/n)(1 - (Y/n))/n}} \right) \approx 1 - \alpha.$$

This gives

A lower bound is

One-sided confidence interval

Example

The Wisconsin Department of Natural Resources (DNR) wished to determine the prevalence, p , of chronic wasting disease (a neurological disease similar to mad cow disease) among its whitetail deer population.

In a particular area of the state, 272 deer were legally killed by hunters in a particular season. A tissue sample from each animal was submitted to the DNR.

Laboratory analysis determined that nine of the deer had chronic wasting disease.

Assume that the 272 harvested deer can, to a good approximation, be regarded as a random sample.

Find a one-sided approximate 95% confidence interval that provides an upper bound for p in that area of Wisconsin.

Difference of two proportions

Suppose Y_1 and Y_2 are independent binomial random variables with n_1, n_2 and p_1, p_2 .

To make a statistical inference about the difference $p_1 - p_2$, we consider $Y_1/n_1 - Y_2/n_2$.

Its mean is

Its variance is

So the CLT gives

Difference of two proportions

Example

Two detergents were tested for their ability to remove stains of a certain type.

An inspector judged the first one to be successful on 63 out of 91 independent trials and the second one to be successful on 42 out of 79 independent trials.

The respective relative frequencies of success are 0.692 and 0.532.

What is an approximate 90% confidence interval for the difference $p_1 - p_2$ of the two detergents?

Exercise

A machine shop manufactures toggle levers.

A lever is flawed if a standard nut cannot be screwed onto the threads.

Let p equal the proportion of flawed toggle levers that the shop manufactures.

There were 24 flawed levers out of a sample of 642 that were selected randomly from the production line.

Find an approximate 95% confidence interval for p .