## Confidence Intervals

**Recall:** A point estimate is a "best guess" function that computes an estimate for a population parameter $\theta$ based on sample data. A point estimate is a statistic. The value of the point estimate depends on the function used and the sample data. For each sample taken, we get to see only the result - one number. This number by itself does not provide information on the precision and reliability of the estimate.

Poll Question 11.1

**Example:** 100 people at a fair guess the number of Jelly Beans in a large glass jar. You, cleverly, record their guesses and compute the average to hand that in as "your" guess in hopes of winning the grand prize of a life-time jelly bean supply.

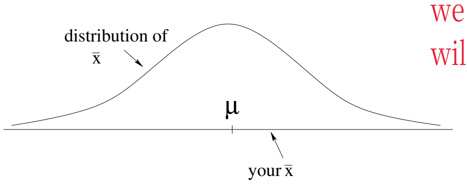
Identify the population parameter and the point estimate(s) in this example. What do you know about the distribution of the point estimate(s)?

true number of JBs in the jar

X_i=guess of i-th person(from 1 to 100), X-bar = your guess

E(X_i)=miu, X-bar ~ Normal(through CLT)

Usually, point estimate functions are chosen, because we hope that their computed values for random samples are close to the true (unknown) population parameter. But how close?



we care about "how far x-bar will be away from miu"

If you know the distribution of $\bar{x}$, you can compute the probability that $\bar{x}$ will fall into any specified interval. Especially, you can compute the probability that $\bar{x}$ will be no further then $c$ from $\mu$.

**Example:** Suppose that there are actually 1250 beans in the jelly bean jar. The guesses of individual people are normally distributed with mean $\mu = 1250$ and standard deviation $\sigma = 70$. You collect data on the guesses of 100 independent people and compute the average $\bar{x}$.

(a) What is the probability that a single persons guess will be within 10 jelly beans of the true answer?

X_i~Normal(1250,70)

P(1240<=X_i<=1260)=0.114

(b) What is the probability that your *average* will be within 10 jelly beans of the true answer?

X-bar ~ Normal(1250,70*70/100)

P(1240<=X-bar<=1260)=0.847

**In General:** Assume that $x_1, \ldots, x_n$ is a random sample from a Normal distribution with (unknown) mean $\mu$ and (known) variance $\sigma^2$. Then $\bar{X}$ has a Normal distribution with mean $\mu$ and variance $\sigma^2/n$. Hence,

make a statement for x bar

this X is RV

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim \text{Normal}(0, 1)$$

$$\Leftrightarrow P(-1.96 \le \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \le 1.96) = 0.95$$

$$\Leftrightarrow P\left(\bar{X} - 1.96\frac{\sigma}{\sqrt{n}} \le \mu \le \bar{X} + 1.96\frac{\sigma}{\sqrt{n}}\right) = 0.95$$

what is different between probability and confidence

**Main Idea:** If $\bar{X}$ is close to $\mu$ with 95% probability, then $\mu$ is also close to $\bar{X}$ with 95% probability. Since we want to estimate $\mu$ from our best guess $\bar{X}$, we can now provide a CONFIDENCE INTERVAL for $\mu$.

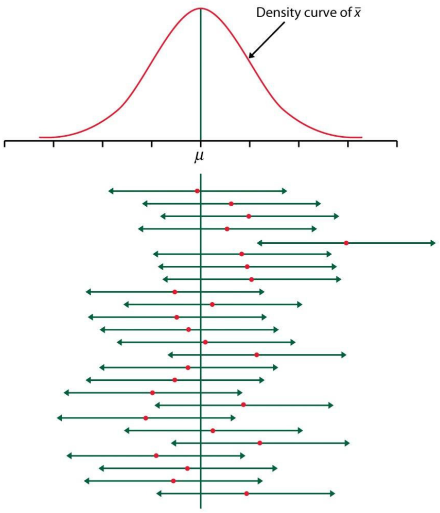
DEFINITION: After observing a random sample $(x_1, \ldots, x_n)$ of size $n$ and computing the sample mean $\bar{x}$, a 95% confidence interval for $\mu$ is

this x is sample

$$\text{CI}_\mu^{95\%} = \left[\bar{x} - 1.96\frac{\sigma}{\sqrt{n}}, \bar{x} + 1.96\frac{\sigma}{\sqrt{n}}\right]$$

INTERPRETING CONFIDENCE INTERVALS

**Caution:** Assume that you collect data $(x_1, \ldots, x_n)$ and then compute a 95% confidence interval for $\mu$ based on $\bar{x}$. It is tempting (but wrong) to conclude that your confidence interval contains the true $\mu$ with probability 0.95. Since the true $\mu$ is a fixed number, it will be contained in *any* interval with probability either 0 or 1. The CONFIDENCE LEVEL (here 95%) rather refers to the generation of the random variables $x_1, \ldots, x_n$ or the sampling process.



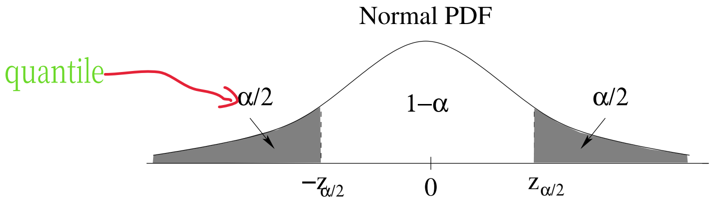
Suppose the fair runs for 20 days and you go each day and observe 100 different people, compute the average of their counts and hand that in as your "best jelly bean guess of the day". On some days you will come close to the true number on the jar (i.e., $\mu$ will be in your confidence interval) but on other days you will not be so close.

On average, for 95% *of samples* your confidence interval will contain the true population parameter $\mu$. If you only play on one day, you do not know whether this day's confidence interval does or does not contain $\mu$, but you are 95% confident that it does.

Poll Question 11.2

## Confidence Levels

To compute confidence intervals for levels other than 95% we have to change the critical $z$-value from 1.96 to another number. You can find the appropriate number in the Normal CDF table.



**Example:** Find the critical values $z_{\alpha/2}$ for the most common choices of confidence levels:

| $1-\alpha$ | 0.9 | 0.95 | 0.99 |
|---|---|---|---|
| $z_{\alpha/2}$ | 1.645 | 1.96 | 2.575 |

Poll Question 11.3

DEFINITION: A **100(1 − α)% confidence interval** for the mean $\mu$ of a normal population when the value of $\sigma$ is known is given by

quantile

confidence level

$$\mathrm{CI}_{\mu}^{(1-\alpha)100\%} = \left[\bar{x} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \bar{x} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right]$$

standard error

parameter

point estimate

## Confidence Level, Precision and Sample Size

Which is better? A 95% confidence interval or a 99% confidence interval? It depends on how much precision is required. Higher confidence level $(1-\alpha)$ means wider confidence intervals (less precision).

What are the factors that influence the width of a confidence interval?

- Confidence level:

  larger confidence level (smaller α) means larger quantile size means less precision

- Sample size:

  larger sample size means smaller standard error, means narrower CI,means more precision

- Standard deviation of $X$:

  larger standard deviation means larger standard error, menas wider CI, means less precision

Poll Question 11.4

in the real cases, the parameter μ varys and standard error are estimated

## Confidence Intervals for Large Samples

If the sample size $n$ is large then the sample mean has approximately a Normal distribution regardless of the population distribution (according to the Central Limit Theorem). However, in most practical applications the population variance $\sigma^2$ is not known. It is still possible to obtain confidence intervals for the mean, by first estimating $\hat{\sigma} = s$ and then using the estimate $s$ in the confidence interval computation.

Proposition: If $n$ is sufficiently large $(n > 40)$, then the standardized variable

$$Z = \frac{\bar{X} - \mu}{S/\sqrt{n}} \sim \text{ Normal}(0, 1)$$

has approximately a standard normal distribution. The $n$ needs to be a little bit larger here than the $(n > 30)$ rule we used for the CLT because of the additional variation introduced through the estimation of $s$.

Proposition: Let $X_1, \ldots, X_n$ be a large random sample from some population with mean $\mu$. Then

from σ to s

$$\text{CI}_{\mu}^{(1-\alpha)100\%} = \left[\bar{x} - z_{\alpha/2}\frac{s}{\sqrt{n}}, \bar{x} + z_{\alpha/2}\frac{s}{\sqrt{n}}\right]$$

is approximately a $100(1 - \alpha)\%$ confidence interval for $\mu$.

**Example:** Suppose the instructor of a large statistics class tries to write an exam of medium difficulty (targeted mean score = 75). Fifty students take the exam and their scores are:

88 93 74 81 ⋯ 98

The mean of these exam scores is 82 with a standard deviation of 18.

(a) Compute a 95% confidence interval for the degree of difficulty of the exam.

μ =the average score for all future students who would take this exam

α =0.05
Z_ α =1.96
n=50
x_bar=82
s=18

Using the formular get[77.01,86.99]

(b) In light of this information, did the exam turn out to be easier, harder or about as hard as the instructor had intended it to be?

easy,because target mean score = 75 which dosen't fall into the CI

Poll Question 11.5

## Confidence Interval for a Population Proportion

Let $p$ denote the (true, but possibly unknown) proportion of "successes" in a population. A random sample of $n$ individuals is selected from the population and let $X$ denote the number of successes in the sample. Then $X$ has approximately a Binomial distribution if the population size is large compared to the sample size. And furthermore, if the sample size $n$ is large as well, then $X$ has approximately a Normal distribution.

Poll Question 11.6

The most commonly used point estimate for $p$ is $\hat{p} = X/n$ the sample proportion of successes.

PROPOSITION: A $100(1-\alpha)\%$ confidence interval for the population proportion $p$ is

$$\mathrm{CI}_p^{(1-\alpha)100\%} = \left[ \frac{\hat{p} + \frac{z_{\alpha/2}^2}{2n} - z_{\alpha/2}\sqrt{\frac{\hat{p}\hat{q}}{n} + \frac{z_{\alpha/2}^2}{4n^2}}}{1 + (z_{\alpha/2}^2)/n}, \frac{\hat{p} + \frac{z_{\alpha/2}^2}{2n} + z_{\alpha/2}\sqrt{\frac{\hat{p}\hat{q}}{n} + \frac{z_{\alpha/2}^2}{4n^2}}}{1 + (z_{\alpha/2}^2)/n} \right]$$

This looks terribly complicated! But if the sample size $n$ is large, then the above formula is approximately equal to

$$\mathrm{CI}_p^{(1-\alpha)100\%} \approx \left[ \hat{p} - z_{\alpha/2}\sqrt{\frac{\hat{p}\hat{q}}{n}}, \hat{p} + z_{\alpha/2}\sqrt{\frac{\hat{p}\hat{q}}{n}} \right]$$

here $\hat{p}$ is your sample proportion of "successes" and $\hat{q} = 1 - \hat{p}$ is the sample proportion of "failures".

The simple confidence interval above can be used in cases where the sample size $n$ is large enough so that $n\hat{p} \geq 10$ and $n\hat{q} \geq 10$.

**Example:** In a Bloomberg survey given in early April 2024 in Michigan, 47% of 4969 registered voters said that they prefer Joe Biden for president the 2024 presidential election and 45% said that they prefer Donald Trump (Other 8%).

Compute an approximate 95% confidence interval for the percentage of voters in Michigan who favor Joe Biden.

$n=4969$

$\hat{p}=0.47$

$\hat{q}=1-0.47$

$Z\_\alpha=1.96$

[0.456,0.484]

## Confidence Intervals for Normal Populations

If we know the population distribution to be Normal, then the sample size may be as small as $n = 2$ for the following to hold:

Proposition: If the population is Normal, and $X_1, \ldots, X_n$ is a random sample from the population, then

$$\bar{X} \sim \text{ Normal}(\mu, \sigma^2/n) \quad \text{or} \quad Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim \text{ Normal}(0, 1)$$

If $\sigma$ is unknown and has to be replaced by the estimated standard deviation $s$, then the distribution changes.

Proposition: Let $\bar{X}$ be the mean of a random sample from a Normal population with mean $\mu$. Then the random variable

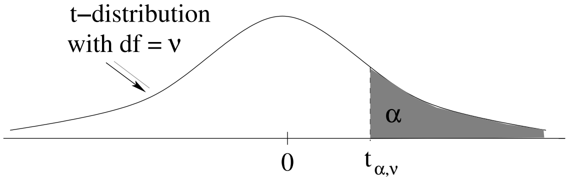
$$T = \frac{\bar{X} - \mu}{S/\sqrt{n}} \sim t_{n-1}$$

has a $t$-distribution with $n - 1$ degrees of freedom.

**Remark:** $t$-distributions look very similar to Normal distributions. They are symmetric, centered at zero and bell shaped. However, they have slightly "thicker" tails than Normal distributions. Values of $t$-distributions are available in tables (in the back of your textbook) or by using your calculator.

**Notation:** Let $t_{\alpha,\nu}$ be the number on the $x$-axis for which the area under the $t$-distribution curve with df $= \nu$ to the right of $t_{\alpha,\nu}$ is equal to $\alpha$. $t_{\alpha,\nu}$ is called a $t$ critical value.

**Example:** Use the table in the back of the book to find $t_{0.05,4}$.



### Poll Question 11.7

Proposition: Let $\bar{x}$ and $s$ be the sample mean and sample standard deviation computed from a random sample of size $n$ from a Normal population with mean $\mu$. Then a $100(1 - \alpha)\%$ confidence interval for $\mu$ is

$$\text{CI}_{\mu}^{(1-\alpha)100\%} = \left[\bar{x} - t_{\alpha/2,n-1}\frac{s}{\sqrt{n}}, \bar{x} + t_{\alpha/2,n-1}\frac{s}{\sqrt{n}}\right]$$

**Example:** A machine is used to fill cans of soft drinks. The machine should be adjusted so that the amount of soft drink in each can is approximately 12 fl oz. Of course it is not possible to fill each can with exactly the same amount, so that the amount of drink per can may be regarded as a Normal random variable with mean 12 and unknown variance $\sigma^2$.

We want to check whether the machine is adjusted correctly and randomly select $n = 10$ cans from the production line and measure precisely the amount of drink in each can:

$$12.08, 11.84, 11.97, \ldots, 12.11$$

Suppose the average $\bar{x}$ of the measurements is 11.95 with sample standard deviation $s = 0.54$.

(a) Compute a 95% confidence interval for $\mu$ the average amount of drink with which the cans get filled.

X_bar=11.95
s=0.54
n=10
α =0.05
t_{ α /2,n-1}=t_(0.025,9)=2.26

(b) Based on your answer above, do you think it is necessary to adjust the machine?

No, we are 95% confident that the machine is already corrected adjusted since 12 is in the CI

**Note of Caution:** the quantity $(\bar{X} - \mu)/(S/\sqrt{n})$ *only* has a $t$-distribution if the population distribution of $X$ is Normal. If the distribution of $X$ is not Normal and the sample size $n$ is small, then a $t$-confidence interval **should not be used**. If it is used, then any interpretation of such a confidence interval might be meaningless!

How do you know in practice, whether your data comes from a Normal distribution? There are procedures to answer this question that you will learn in a subsequent statistics class (such as Math 161B).

- probability plots or quantile plots
- hypothesis tests (Kolmogorov-Smirnov, Anderson-Darling etc.)

In practice, one would employ one of these methods first, before using a $t$-distribution based procedure.

**Question:** What do you do if you find that your data is not Normal and your sample is not large enough to use a Normal approximation? Can you still find a confidence interval for the mean? Yes! But the methods you would have to use go beyond the scope of this class.

- Bootstrap
- Wilcoxon signed-rank procedure if population distribution is symmetric