Applied Statistical Decision Making
Confidence Intervals

Monia Ranalli
monia.ranalli@uniroma2.it

Objectives

- understand the difference between **point and interval estimates**
- calculate and interpret **confidence intervals for population proportions**
- know the meaning of **margin of error** and its use
- determine the required **sample size** for different experimental setting
- know when and how to use t-**interval to estimate the population mean**
- compute **sample sizes for estimating the population mean**

Introduction to Inferences

Based on sample observed data, statistical inference applies the concepts of probability to give conclusions about the population of interest

Two types of statistical inferences

- **Estimation**: use information from the sample to estimate (or predict) the parameter of interest

  Example: using the result of a poll about the president's current approval rating to estimate (or predict) his or her true current approval rating nationwide.

- **Statistical Test**: use information from the sample to determine whether a certain statement about the parameter of interest is true

  Example: the democrats claim that President Barack Obama's current approval rating is more than 85%. We want to determine whether that statement is supported by the poll data.

Properties of 'Good' Estimators

- **Property 1**. A good estimator has a sampling distribution that is centered at the parameter of interest, i.e. the mean of the distribution equals the true value of the parameter. An estimator with this property is **unbiased**.

  The sample mean is an unbiased estimator of the population mean. The sample proportion is an unbiased estimator of the population proportion.

- **Property 2**. A good estimator has the **smallest standard error** compared to other estimators.

  This means it tends to fall closer than other estimates to the parameter. For example, the sample mean has a smaller standard error than the sample median when estimating the population mean of a normal distribution. In the normal distribution the mean and median are essentially the same. However, the standard error of the median is about 1.25 times that of the standard error of the mean. The standard error of the mean is $\sqrt{\sigma/n}$*. Therefore in a normal distribution, the se(median) is about 1.25 times* $\sqrt{\sigma/n}$*. This is why the mean is a better estimator than the median when the data is normal (or approximately normal).*

Example

Population (1 = speak a second language)

X	f_k
0	0.80
1	0.20

Sample Space for n=2

X_1, X_2	$p(X_1, X_2)$	$\hat{\pi}$
0,0	0.8×0.8	0
0,1	0.8×0.2	0.5
1,0	0.2×0.8	0.5
1,1	0.2×0.2	1

$\hat{\pi}$	p	$\hat{\pi} * p(\hat{\pi})$	$p(\hat{\pi}) * (\hat{\pi} - 0.2)^2$
0	0.64	0	0.64*0.04 = 0.0256
0.5	0.320	0.16	0.32*0.09 = 0.0288
1	0.04	0.04	0.04*0.64 = 0.0256
	1	0.20=E($\hat{\pi}$)	0.0800=Var($\hat{\pi}$)

X_1	$p(X_1)$	$X_1 * p(X_1)$	$p(X_1) * (X_1 - 0.2)^2$
0	0.8	0	0.8*0.04 = 0.032
1	0.2	0.2	0.2*0.64 = 0.128
	1	0.20=E(X_1)	0.160=Var(X_1)

$\hat{\pi}$ and X_1 are both unbiased, but $\hat{\pi}$ is better than X_1 because it has a smaller variance $\Rightarrow$ smaller standard deviation $\Rightarrow$ smaller standard error

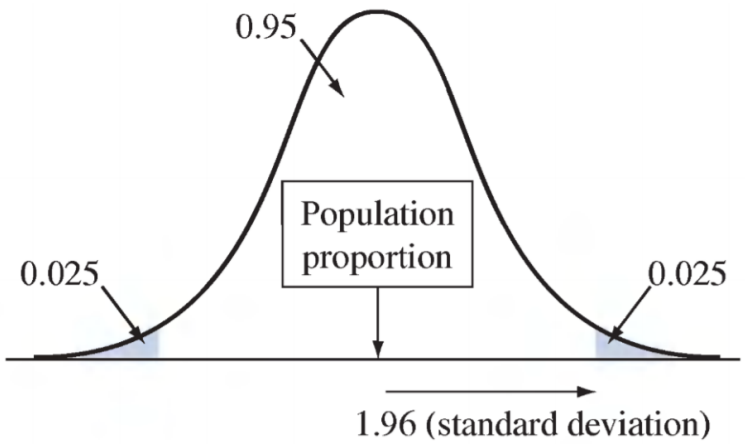
Confidence Intervals and its Logic

- **Confidence interval**: it is an interval containing the most believable values for a parameter
- **Confidence level**: it is the probability that this method produces an interval that contains the parameter. This is a number chosen to be close to 1, most commonly 0.95
- **General construction of a confidence interval**: **sample statistic** $\pm$ **margin of error**
- **Margin of error**. It consists of two pieces: the standard error of the sample statistic and some multiplier of this standard error depending on the confidence level.
- **Interpretation**: "We are 'some level of percent confident' that the 'population of interest' is from 'lower bound to upper bound".

Inferences on the Binomial Parameter π

To construct a **confidence interval for a population proportion**, the sampling distribution of a sample proportion

- gives the possible values for the sample proportion and their probabilities
- is approximately a normal distribution for large random samples by the CLT
- has mean equal to the population proportion



With probability 0.95, the sample proportion falls within about 1.96 standard errors of the population proportion. The distance of 1.96 standard errors is the margin of error in calculating a 95% confidence interval for the population proportion.

Constructing a Confidence Interval to Estimate a Population Proportion

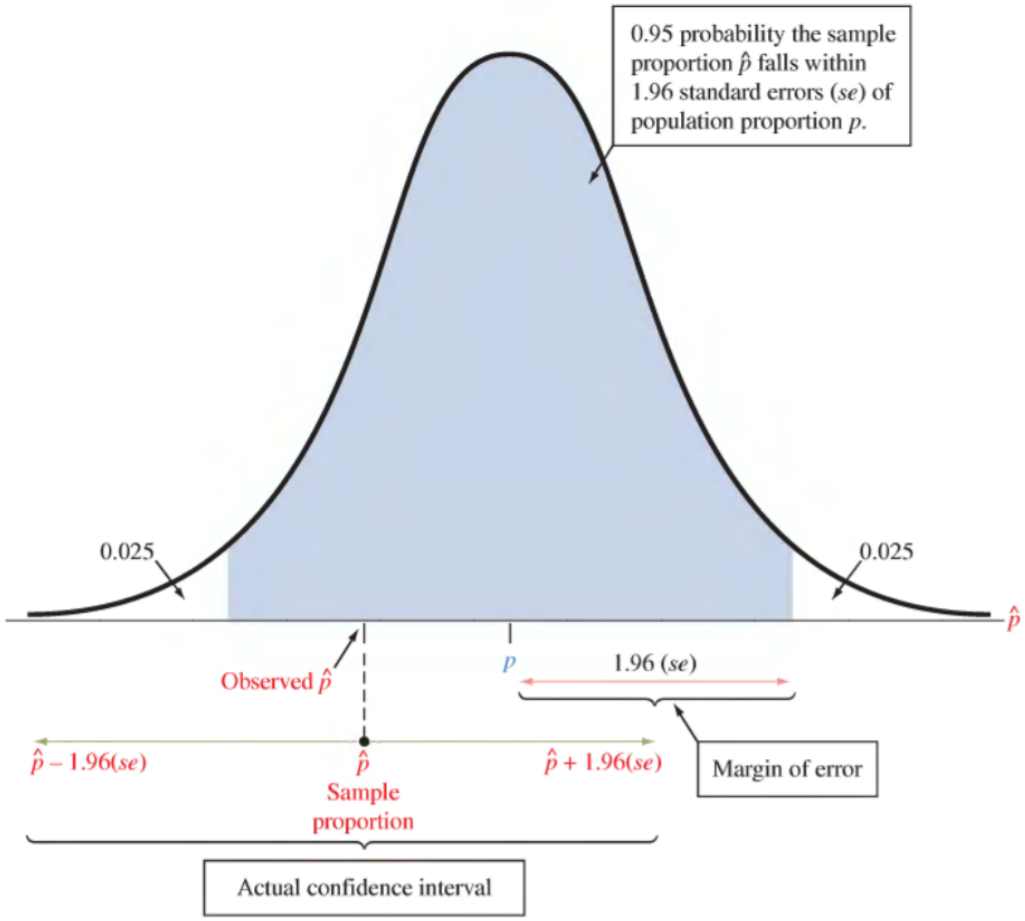
- **Population proportion**: π
- **Point estimate of the population proportion**: the sample proportion, $\hat{\pi}$
- **Assumptions needed to check** before applying CLT: $n\hat{\pi} \geq 15$ **AND** $n(1-\hat{\pi}) \geq 15$

$$\Downarrow$$

$(1-\alpha)100\%$ confidence interval for π is: $\hat{\pi} \pm z_{\alpha/2}\sqrt{\dfrac{\hat{\pi}(1-\hat{\pi})}{n}}$,
where $z_{\alpha/2}$ represents a z-value with $\alpha/2$ area to the right of it

NOTE: Previously we used the population proportion π instead of $\hat{\pi}$ to compute the standard error of a sample proportion or to check the assumptions of the CLT. However, the aim is to estimate that value (π), that is, the population proportion $\Rightarrow$ we need to adjust by using the estimate $\hat{\pi}$

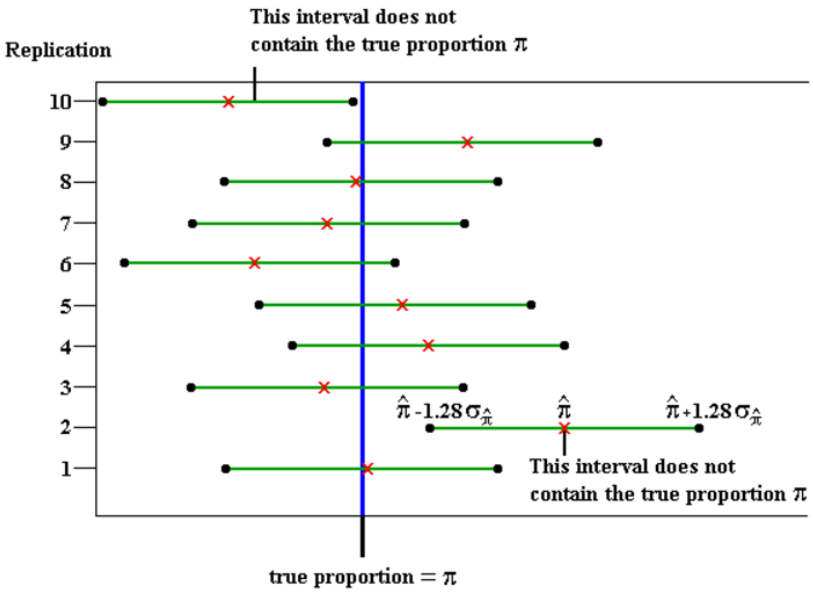
Example



A survey asked to a sample of 1900 persons: "do you believe in hell?"; 1387 responded yes. The **point estimate** for the proportion of adult Americans who would respond yes equals 0.73. The **margin of error** (with a probability of 95%) is $1.96 * se = 1.96 * 0.444/\sqrt{1900} = 0.02$. The **interval estimate** is $[0.71, 0.75]$.

More on the Interpretation of a Confidence Interval

For each replication, 30 students are sampled and it has been asked them whether they are Democrats. For each replicate an 80% CI has been computed. In this set of 10 replications, exactly 8 out of 10 intervals contain the parameter. Due to the small number of replications (only 10), it is quite possible that 9 out of 10 or 7 out of 10 contain the true parameter. On the other hand, if the number of replicates is 10.000 (more in general, a large number), the percentage that contains the true proportions will be very close to 80%.



If random samples of size n are repeatedly drawn from the population where the proportion of success in the population is π and the confidence interval $\hat{\pi} \pm z_{\alpha/2}\sqrt{\hat{\pi}(1-\hat{\pi})/n}$ is calculated each time, 100(1 -)% of the intervals are expected to would contain the true parameter, π.

Sample Size Computation for Population Proportion Confidence Interval

- **Margin of Error**: it is half of the width of the confidence interval
$$E = z_{\alpha/2}\sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n}}$$
- **Confidence and precision**: the higher the confidence level, the wider the width (or equivalently, half width) of the interval and thus the poorer the precision.
- To get a **smaller margin of error** (i.e. narrower intervals), the sample size should be increased. Or, the margin of error will also be smaller if a 90% confidence interval instead of a 95% confidence interval is computed.

Example

During a referendum, in a sample of $n = 100$ exit polls, we found a proportion of 0.55 YES. What can we conclude?

First of all, we calculate confidence intervals for the proportion

- 95% CI: $0.55 \pm 1.9600 * 0.497/\sqrt{100} = 0.55 \pm 0.097$
  $\Rightarrow 0.453 < \pi < 0.647$
- 99% CI: $0.55 \pm 2.5758 * 0.497/\sqrt{100} = 0.55 \pm 0.128$
  $\Rightarrow 0.422 < \pi < 0.678$

We observe that for π are plausible values of less than 0.5, so we can not yet conclude anything about the outcome of the referendum.

Now suppose to have $n = 1000$ exit polls that confirm an estimate of 0.55 for the YES. What can we conclude?

We observe that the confidence intervals now are

- 95% CI: $0.55 \pm 1.9600 * 0.497/\sqrt{100} = 0.55 \pm 0.031$
  $\Rightarrow 0.519 < \pi < 0.581$
- 99% CI: $0.55 \pm 2.5758 * 0.497/\sqrt{100} = 0.55 \pm 0.040$
  $\Rightarrow 0.510 < \pi < 0.590$

and we can therefore conclude in favor of the victory of the YES.

Example

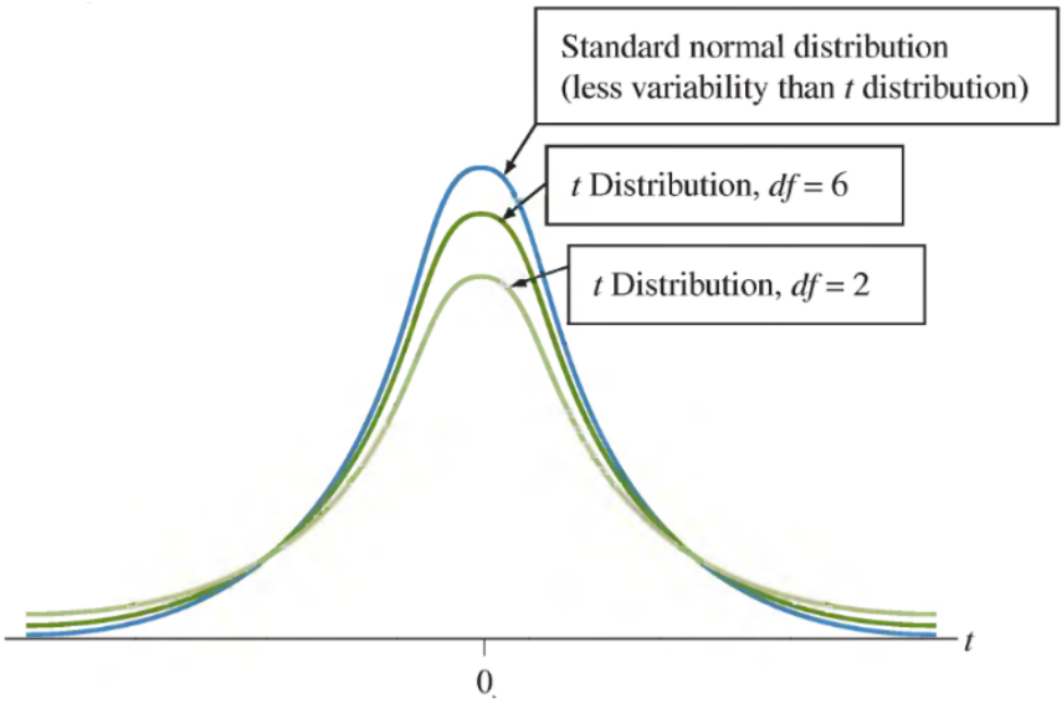
A random sample of 1500 U.S. adults is taken. They are asked whether they approve or disapprove of the current president's performance so far (i.e. an approval rating). Of the 1500 surveyed, 660 respond with "approve". Calculate a 95% confidence interval for the overall approval rating of the the president.

1. **Check conditions**. $\hat{\pi}n$ and $(1-\hat{\pi})$ are both greater than 15. $\Rightarrow$ sampling distribution is well approximated by a normal distribution
2. **Find the $z_{\alpha/2}$ multiplier** for the level of confidence (i.e. 1.96)
3. **Compute the 95% CI**: $0.44 \pm 1.96\sqrt{\frac{0.44 * (1-0.44)}{1500}}$
   $\Rightarrow 0.415 < \pi < 0.465$
4. **Provide an interpretation of the interval:** "We are 95% confident that the overall U.S. adult approval rating for the current president is from 41.5% to 46.5%." Alternatively, "The current U.S. approval rating for the president is 44% with a 95% margin of error of 2.5%."

Constructing a Confidence Interval to Estimate a Population Mean I

- **Population mean**: μ
- **Point estimate of the population mean**: the sample mean, $\bar{x}$
- The **exact standard error of the sample mean**: $\sigma/\sqrt{n}$
- **In practice**, we estimate σ by the **sample standard deviation**, s
- **For large** n from any population and also for every n from an underlying population that is normal, the confidence interval for the population mean is $\bar{x} \pm z_{\alpha/2}\dfrac{\sigma}{\sqrt{n}}$
- In practice, the **population standard deviation is not known**
- Substituting the sample standard deviation s for σ to get $se = s/\sqrt{n}$ introduces **extra error**
- To account for this increased error, **the z-score is replaced** by a slightly larger score, the **t-score**

Properties of the t Distribution



- The t-distribution is **bell shaped** and **symmetric** about 0 (=mean)
- The probabilities depend on the **degrees of freedom**, $df = n - 1$
- The t-distribution has **thicker tails** than the standard normal distribution, i.e., it is **more spread out**
- As $n \to \infty$, $t \to Z$ (by CLT)

Constructing a Confidence Interval to Estimate a Population Mean II

- When the standard deviation of the population is unknown, a 95% confidence interval for the population mean μ is

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

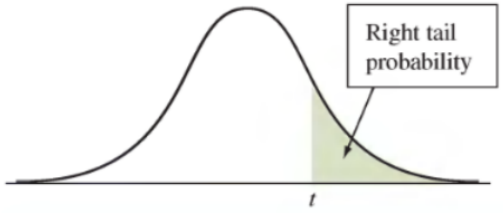


Table B *t* Distribution Critical Values

	Confidence Level					
	80%	90%	95%	98%	99%	99.8%
	Right-Tail Probability					
df	$t_{.100}$	$t_{.050}$	$t_{.025}$	$t_{.010}$	$t_{.005}$	$t_{.001}$
1	3.078	6.314	12.706	31.821	63.656	318.289
2	1.886	2.920	4.303	6.965	9.925	22.328
3	1.638	2.353	3.182	4.541	5.841	10.214
4	1.533	2.132	2.776	3.747	4.604	7.173
5	1.476	2.015	2.571	3.365	4.032	5.894
6	1.440	1.943	2.447	3.143	3.707	5.208

Constructing a Confidence Interval to Estimate a Population Mean III

- A basic **assumption** of the confidence interval using the t-distribution is that the **population distribution is normal** $\rightarrow$ t-distribution is a **robust method** in terms of the normality assumption
- Many variables have distributions that are far from normal. How problematic is it?
- For **large random samples**, it is not problematic because of the Central Limit Theorem
- **If n is small**, confidence intervals using t-**scores usually work quite well** except for when extreme outliers are present. $\rightarrow$ The **method is robust**

Example

A study of 7 American adults from an SRS yields an average height of 67.2 inches and a standard deviation of 3.9 inches. Assuming the heights are normally distributed, 95% and 99% confidence intervals for the average height of all American adults (μ) are:

- 95% CI: $67.2 \pm 2.447 * 3.9/\sqrt{7} = 67.2 \pm 3.607$
  $\Rightarrow 63.593 < \mu < 70.807$
  "We are 95% confident that the average height of all American adults is between 63.6 and 70.8 inches"
- 99% CI: $67.2 \pm 2.447 * 3.707/\sqrt{7} = 67.2 \pm 5.465$
  $\Rightarrow 61.735 < \mu < 72.665$
  "We are 99% confident that the average height of all American adults is between 61.7 and 72.7 inches"

General comments

- **Confidence and precision**: the higher the confidence level, the wider the width of the interval and thus less precision
- **Confidence intervals depend on:**
  - level of confidence
  - sample size
  - standard error

Interval width	$(1-\alpha)\%$	s	n
$\uparrow$	$\uparrow$	$\uparrow$	$\downarrow$
$\downarrow$	$\downarrow$	$\downarrow$	$\uparrow$

Sample Size Computation for the Population Mean

- **Margin of Error**: it is half width of the interval,

$$E = \frac{t_{\alpha/2} s}{\sqrt{n}}$$

- **Sample Size Computation for Population Mean**: given the confidence level and the half width of the interval, it is possible to determine the sample size to yield an interval with that confidence level and with a half width not more than the specified one,

$$n^\star = \left(\frac{z_{\alpha/2}\sigma}{E}\right)^2,$$

rounded up to the next whole integer.