Topic 4: Sampling Distribution and Large-Sample Estimation

Optional Reading: Chapter 7 and 8

Xiner Zhou

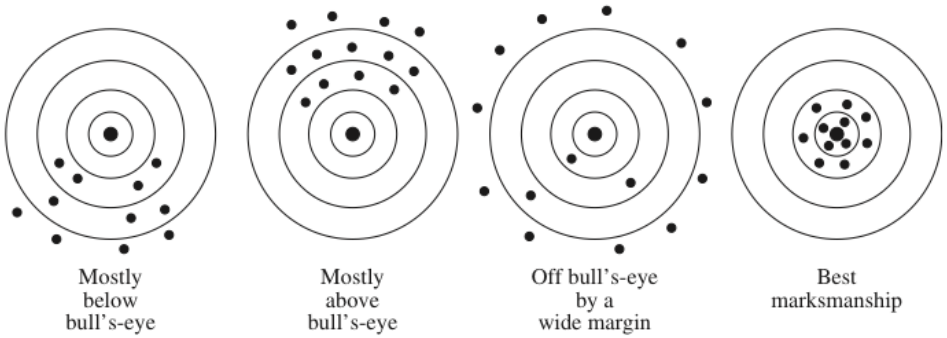
Department of Statistics

University of California, Davis

- Central Limit Theorem for Sampling Distribution
- Large-Sample estimation (point estimation and confidence interval)
  - One population mean
  - Difference between two population means
  - One population proportion
  - Difference between two population proportions

Point Estimation

Rationale for point estimation:



Bull's eye shooting

- Parameter of interest = bull's eye
- Your aim is to fire bullet hit the bull's eye

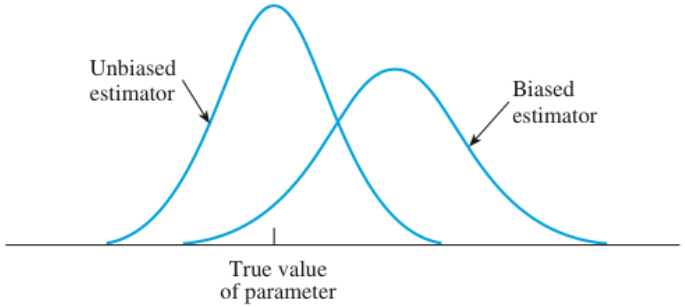
Which one has the best shot?

Reasonable people would prefer the shooter who do not consistently miss the target by a wide margin, and shots are cluster closely around the target.

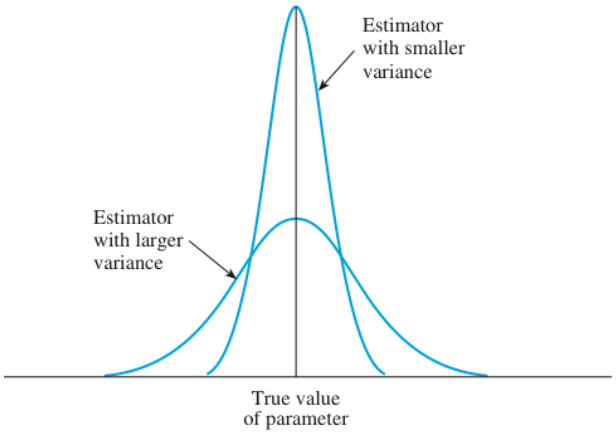
➡ Same principle applies to point estimation

Same principle applies to point estimation

1. Sampling distribution of appropriate point estimator should be centered at the true value of the parameter: unbiased estimator
   - which we don't know, so we have unknown target, harder than bull's eye shooting



2. The spread of the sampling distribution should be as small as possible
   - More accurate in your shooting = greater confidence and certainty about your decision



How to estimate a population mean?

To estimate the mean μ of a population
Suppose we observe sample: $X_1, \ldots X_n$.

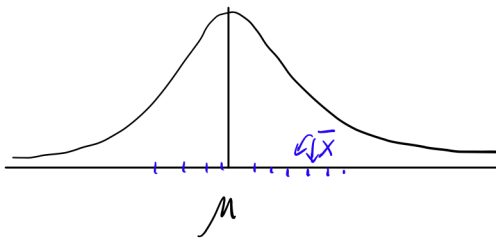
Then, the point estimator for the mean μ is the sample mean $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$

Due to random sampling, the sample mean $\bar{X}$ will not be exactly the same as μ

If we report just $\bar{X}$, which is just the best guess for μ
it does not fully specify where μ could be.

By CLT, we know the sampling distribution of the sample mean is:

$$\bar{X} \sim N\left(\mu_1, \left(\frac{\sigma}{\sqrt{n}}\right)^2\right)$$

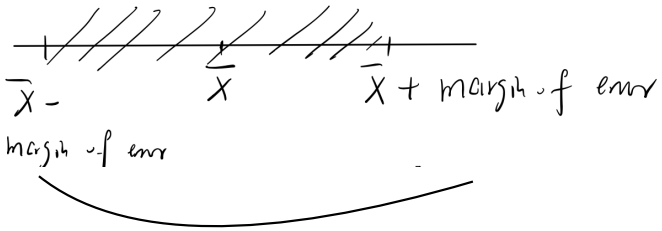


Due to the fact of Normal distribution or the 68-95-99.7% rule:

$$P\left(|\bar{X} - \mu| < 1.96\frac{\sigma}{\sqrt{n}}\right) \approx 95\,\%$$

With extremely high probability (95%), the estimator we report, i.e. the sample mean, will only deviate from truth μ by $\pm 1.96\frac{\sigma}{\sqrt{n}}$.

We call $1.96\frac{\sigma}{\sqrt{n}}$ 95% margin of error



Provide a range where truth may lie
It's very unlikely that the truth will exceed the range

An environmentalist is conducting a study of the polar bear, a species found in and around the Arctic Ocean. Their range is limited by the availability of sea ice, which they use as a platform to hunt seals, the mainstay of their diet. The destruction of its habitat on the Arctic ice, which has been attributed to global warming, threatens the bear's survival as a species; it may become extinct within the century.[1] A random sample of $n = 50$ polar bears produced an average weight of 980 pounds with a standard deviation of 105 pounds. Use this information to estimate the average weight of all Arctic polar bears.

Solution: weight of polar bear ~ some unknow distribution with mean μ

population

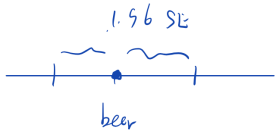
data: $n = 50$ $\bar{x} = 980$ $s = 105$

$\Rightarrow$ point estimate for average weight of all Arctic polar bears

is, $\bar{x} = 980$

with 95% margin of error:

$$1.96\,SE = 1.96\left(\frac{s}{\sqrt{n}}\right) = 1.96\left(\frac{105}{\sqrt{50}}\right) = 29.1$$



We are confident that the difference between estimate 980 and true μ is within $\pm$ 29.1,

true μ could be as small as $980 - 29 = 951$

and as large as $980 + 29 = 1009$

How to estimate a population proportion?

To estimate the proportion p of a Binomial population
Suppose we observe sample: X= total number of “successes” or “events"

Then, the point estimator for the proportion p is the sample proportion

$$\hat{p} = \frac{X}{n}$$

Due to random sampling, the sample proportion $\hat{p}$ will not be exactly the same as the population proportion p

If we report just $\hat{p}$, which is just the best guess for p
it does not fully specify where p could be.

By CLT, we know the sampling distribution of the sample proportion is:

$$\hat{p} \sim N(p, (\sqrt{\frac{p(1-p)}{n}})^2)$$

95% margin of error: $1.96\sqrt{\frac{p(1-p)}{n}} \approx 1.96\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$

In addition to the average weight of the Arctic polar bear, the environmentalist from Example 8.4 is also interested in the opinions of adults on the subject of global warming. In particular, he wants to estimate the proportion of adults who think that global warming is a very serious problem. In a random sample of $n = 100$ adults, 73% of the sample indicated that global warming is a very serious problem. Estimate the true population proportion of adults who believe that global warming is a very serious problem, and find the margin of error for the estimate.

Solution: point estimate for p: $\hat{p}$ = sample prop. = 0.73

margin of error: $1.96\,SE = 1.96\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$

$= 1.96\sqrt{\frac{0.73 \times 0.27}{100}}$

$= 0.09$

We are confident that the difference between estimate 0.73 and true p is within ± 0.09,

true p could be as small as $0.73 - 0.09 = 0.64$

and as large as $0.73 + 0.09 = 0.82$

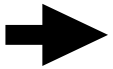
Confidence Interval

Rationale for constructing a Confidence Interval:

The point estimate gives a single number targeting the parameter of interest.
It will not be exactly equal to the true parameter.

The margin of error quantify how large can the difference between point estimate and true parameter be.

If the data provide more information, then we're more certain about our point estimate, and the margin of error will be smaller;
otherwise, the margin of error will be larger, reflecting more uncertainty.

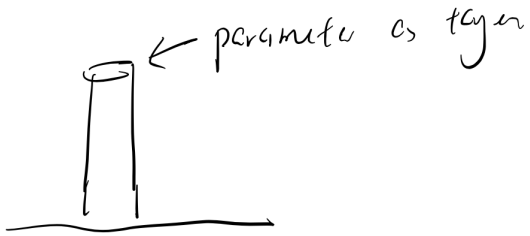


Confidence interval can be thought as a combination of both point estimate + margin of error,
It gives a range of values that the true parameter could take.

Like lariat roping:

Parameter = Fence post ← confidence interval

Interval estimate = Lariat



You hope to include the fence post by the rope
= confidence interval: you hope to include the true parameter by the interval you calculate

Proportion of times the rope does include the post, if you repeatedly throw the rope is your “success rate”
= confidence coefficient $1-\alpha$ (α missing %)

- If too small: often miss the target
- Usually: 0.9, 0.95, 0.99

General formula for $(1-\alpha)100\,\%$ confidence interval

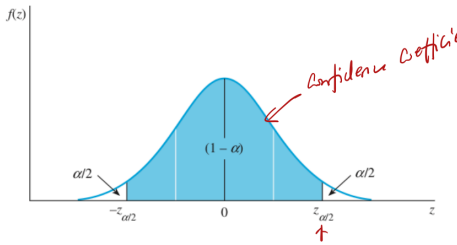
(point estimator) $\pm$ $z_{\alpha/2}$ $\times$ (standard error of the estimator)

$z_{\alpha/2}$: the value of Z that has tail area $\frac{\alpha}{2}$ to its right

= margin of error !

= [LCL , UCL]

LCL: lower confidence limit UCL: upper confidence limit



Confidence coefficient: area in the middle

$(1-\frac{\alpha}{2})$th percentile of standard normal

Values of z Commonly Used for Confidence Intervals

Confidence Coefficient, $(1-\alpha)$	α	$\alpha/2$	$z_{\alpha/2}$
.90	.10	.05	1.645
.95	.05	.025	1.96
.98	.02	.01	2.33
.99	.01	.005	2.58

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

$$\bar{X} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

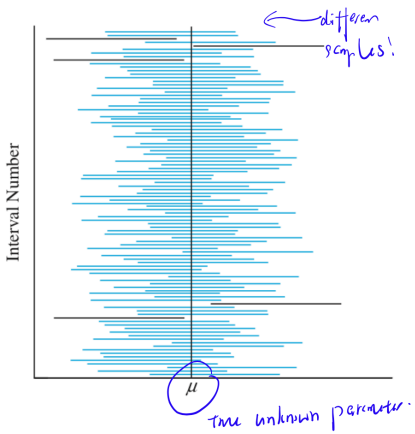
If σ is unknown, replace with sample standard deviation $S = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2}$

$(1-\alpha)100\,\%$ confidence interval for population proportion p

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

Interpretation the confidence interval

What does “95% confidence interval” mean?



You can not be absolutely sure that, in any one particular experiment, the confidence interval contains the true parameter.

You will never know whether this particular interval is one of those “missed” or “covered” ones.

➡ ~~There is 95% chance that the true parameter lies in the confidence interval.~~

➡ Correct interpretation:
If you repeatedly conduct the same experiment for many times under the same condition, and every time you construct a confidence interval, then you can expect that, 95% of those confidence intervals cover the true parameter.

A dietician selected a random sample of $n = 50$ male adults and found that their average daily intake of dairy products was $\bar{x} = 756$ grams per day with a standard deviation of $s = 35$ grams per day. Use this sample information to construct a 95% confidence interval for the mean daily intake of dairy products for men.

Solution: $\bar{x} \pm z_{\alpha/2}\left(\frac{s}{\sqrt{n}}\right)$

$1-\alpha = 0.95 \Rightarrow \alpha = 0.05 \Rightarrow z_{\alpha/2} = 1.96$

$756 \pm 1.96\left(\frac{35}{\sqrt{50}}\right) = 756 \pm 9.7 = [746.3, 765.7]$

A random sample of 985 "likely" voters—those who are likely to vote ~~in the upcoming~~ election—were polled during a phone-athon conducted by the Republican Party. Of those surveyed, 592 indicated that they intended to vote for the ~~Republican~~ candidate in the upcoming election. Construct a 90% confidence interval for p, the proportion of likely voters in the population who intend to vote for the Republican candidate. Based on this information, can you conclude that the candidate will win the election?

biased sample

⇒ 90% confidence

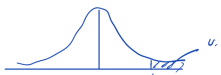
✗

Solution: $\hat{p} \pm z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$

$\hat{p} = \frac{x}{n} = \frac{592}{985} = 0.601$

$\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = \sqrt{\frac{0.601 \times 0.399}{985}} = 0.016$

$z_{\alpha/2}$:

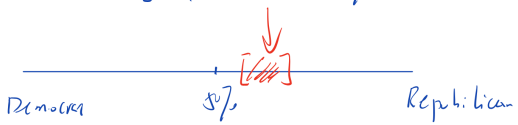


$z_{\alpha/2} = 1.645$

⇒ 90% C.I. : $0.601 \pm 1.645 \times 0.016$

$= 0.601 \pm 0.026$

$= [57.5\%, 62.7\%]$

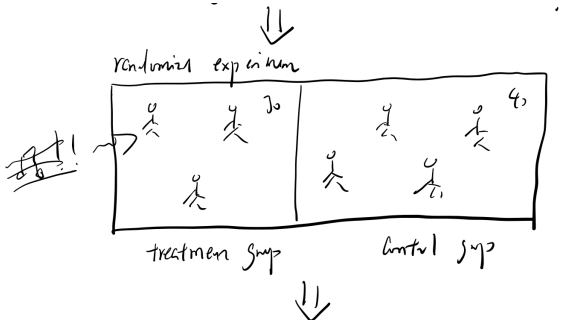


If : majority win
sample is not biased ?
response accurately reflect actual voting behavior ?

⇒ 90% confidence that the Republican candidate will win.!

Noise and Stress

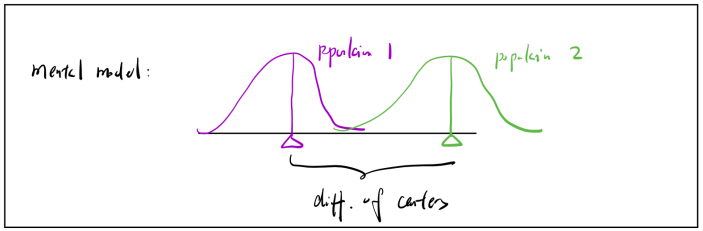
What's the effect of noise as a source of stress on the ability to perform simple tasks?



Record "time to finish the task"

	Control	Experimental
n	30	40
$\bar{x}$	15 minutes	23 minutes
s	4 minutes	10 minutes

Many questions are comparison between two populations, specifically, looking at the difference between two population means. E.g. is the new vaccine more effective than the old one? ….



	Population 1	Population 2
Mean	μ_1	μ_2
Variance	σ_1^2	σ_2^2

To answer these type of questions, we draw random samples from both populations

	Sample 1	Sample 2
Mean	$\bar{x}_1$	$\bar{x}_2$
Variance	s_1^2	s_2^2
Sample Size	n_1	n_2

We then use the sample to estimate the difference between two population means.

Estimation for Difference between two population means

To estimate the difference between two population means $\mu_1 - \mu_2$

Suppose we observe sample:
$X_{1,1}, \ldots X_{1,n_1}$ from population 1
$X_{2,1}, \ldots X_{2,n_2}$ from population 2

Then:

1. the point estimator for $\mu_1 - \mu_2$: difference of sample means
$\bar{X}_1 - \bar{X}_2$

2. Sampling distribution of $\bar{X}_1 - \bar{X}_2$ is
$$\bar{X}_1 - \bar{X}_2 \sim N\left(\mu_1 - \mu_2, \frac{\sigma_1^2}{n_1} + \frac{{\sigma_2}^2}{n_2}\right)$$
   - Standard error = $\sqrt{\frac{\sigma_1^2}{n_1} + \frac{{\sigma_2}^2}{n_2}}$
   - If σ_1, σ_2 unknown, replace with sample standard deviations S_1, S_2

3. 95% margin of error: $1.96\sqrt{\frac{\sigma_1^2}{n_1} + \frac{{\sigma_2}^2}{n_2}}$

4. $(1-\alpha)100\,\%$ confidence interval
$$\bar{X}_1 - \bar{X}_2 \pm z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{{\sigma_2}^2}{n_2}}$$

Noise and Stress

What's the effect of noise as a source of stress on the ability to perform simple tasks?

Record "time to finish the task"

	Control	Experimental
n	30	40
$\bar{x}$	15 minutes	23 minutes
s	4 minutes	10 minutes

Solution: 1. What's the estimated difference in means completion time for the two groups?

$\bar{x}_1 - \bar{x}_2 = 23 - 15 = 8$

2. 99% Margin of error ?

$\pm z_{\alpha/2} \times SE = \pm 2.58 \sqrt{\frac{10^2}{40} + \frac{4^2}{30}}$

$= \pm 4.49$

$\Rightarrow$ 99% confidence that the time diff. $(\mu_1 - \mu_2)$ is above or below the estimated diff $(\bar{x}_1 - \bar{x}_2)$ by or more 4.49

3. 99% C.I. ?

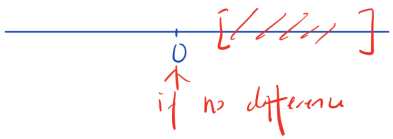
$(\bar{x}_1 - \bar{x}_2) \pm z_{\alpha/2} SE$

$= 8 \pm 4.49$

$= [3.51, 12.49]$

$\Rightarrow$ The diff. in completion time between noise vs no-noise groups is estimated to lie between LCL = 3.51 and UCL = 12.49 seconds.

4. Based on the C.I., is there sufficient evidence to indicate a real difference in the average time to complete for the two groups?



Decision:

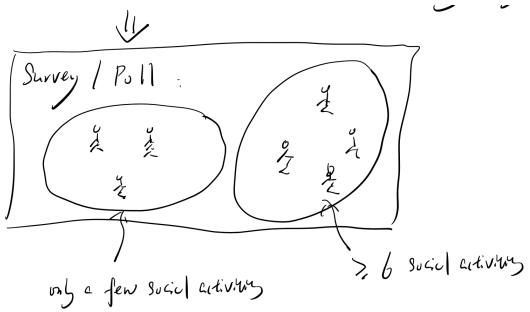
Since 0 is not one of the possible values for $(\mu_1 - \mu_2)$ as indicated by 99% C.I.,

it is not likely that the mean completion times for noise vs non-noise groups are the same.

In fact, since the entire 99% C.I. lies on the positive half of the line, we're fairly certain that noise makes the average completion time for a simple task longer.

Social activities vs Getting colds

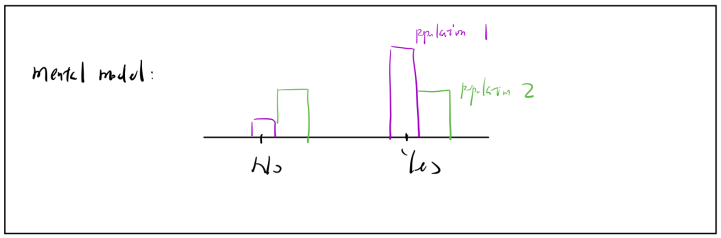
Do people with more social activities less likely or more likely to get colds?



	Few Social Outlets	Many Social Outlets
Sample Size	96	105
Percent with Colds	62%	35%

Many questions are comparison between two population proportions.

- The proportion of defective items manufactured in two production lines
- The proportion of male and female voters who favor an equal rights amendment



To answer these type of questions, we draw random samples from both populations

We then use the sample to estimate the difference between two population means.

Estimation for Difference between two population proportions

To estimate the difference between two population proportions $p_1 - p_2$

Suppose we observe sample:
n_1 sample from population 1: X_1= total number of subjects with “success” or “event” happened
n_2 sample from population 2: X_2= total number of subjects with “success” or “event” happened

Then:

1. the point estimator for $p_1 - p_2$: difference of sample proportions

$$\hat{p}_1 - \hat{p}_2 = \frac{X_1}{n_1} - \frac{X_2}{n_2}$$

2. Sampling distribution of $\hat{p}_1 - \hat{p}_2$ is

$$\hat{p}_1 - \hat{p}_2 \sim N\left(p_1 - p_2, \frac{\hat{p}_1\left(1-\hat{p}_1\right)}{n_1} + \frac{\hat{p}_2\left(1-\hat{p}_2\right)}{n_2}\right)$$

- Standard error = $\sqrt{\frac{\hat{p}_1\left(1-\hat{p}_1\right)}{n_1} + \frac{\hat{p}_2\left(1-\hat{p}_2\right)}{n_2}}$

3. 95% margin of error: $\pm 1.96\sqrt{\frac{\hat{p}_1\left(1-\hat{p}_1\right)}{n_1} + \frac{\hat{p}_2\left(1-\hat{p}_2\right)}{n_2}}$

4. $(1-\alpha)100\,\%$ confidence interval

$$\hat{p}_1 - \hat{p}_2 \pm z_{\alpha/2}\sqrt{\frac{\hat{p}_1\left(1-\hat{p}_1\right)}{n_1} + \frac{\hat{p}_2\left(1-\hat{p}_2\right)}{n_2}}$$

Social activities vs Getting colds

Do people with more social activities less likely or more likely to get colds?

	Few Social Outlets	Many Social Outlets
Sample Size	96	105
Percent with Colds	62%	35%

Solution: 1. What's the estimated difference in proportions (many social − few)?

$\hat{p}_1 - \hat{p}_2 = 35\% - 62\% = -27\%$.

2. 99% Margin of error?

$\pm z_{\alpha} \times SE = \pm 2.58 \sqrt{\frac{0.35 \times 0.65}{105} + \frac{0.62 \times 0.38}{96}}$

$= 17.5\%$

⇒ 99% confidence that the true diff $(p_1 - p_2)$ is above or below the estimated diff $(\hat{p}_1 - \hat{p}_2)$ by as much 17.5%

3. 99% C.I.?

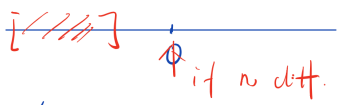
$(\hat{p}_1 - \hat{p}_2) \pm z_{\alpha} SE$

$= -27\% \pm 17.5\%$

$= [-44.5\%, -9.5\%]$

⇒ The diff. in proportion of getting colds between two groups is estimated to lie between UCL = −44.5% and LCL = −9.5%.

4. Does there appear to be a difference in the population proportions for the two groups?



Conclusion/Decision:

Since 0 is not one of the possible values for $(p_1 - p_2)$ as indicated by 97% C.I.,
it is not likely that the population proportions for the two groups are the same.

In fact, since the entire 97% C.I. lies on the negative side of the line, we're fairly certain that people with ≥ 6 social activities are less likely to get colds, than people with few social activities.

5. Open question:

Does your intuition tell you: more contact with people $\Rightarrow$ more colds?

If so, the data show the opposite effect.

How can you explain this unexpected finding?