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

Descriptive statistics

Probability

Random variable and its distribution

Where we have been

Inference problems:

- Decision making
  - The new vaccine is more effective than an old one?
  - A home-buyer wants to estiamte the market price for a house before putting out an offer
- Prediction
  - A financial analyst need to predict the behavior of stock market
  - Political scientists need to predict the outcome of an election

....

It’s the job of statistics to make objective suggestions for decision making and predictions, based on data, i.e. “let the data speak!”

First, we need to have a probability model for each problem, e.g. Binomial experiment

- The new vaccine is more effective than an old one?

p1: probability of getting sick who received new vaccine
p2: probability of getting sick who received the old vaccine

Sample from the population who received the new vaccine ~ Binomial(n1, p1)
Sample from the population who received the oldvaccine ~ Binomial(n1, p2)

Question: p1 <p2 ? Hypothesis testing problem

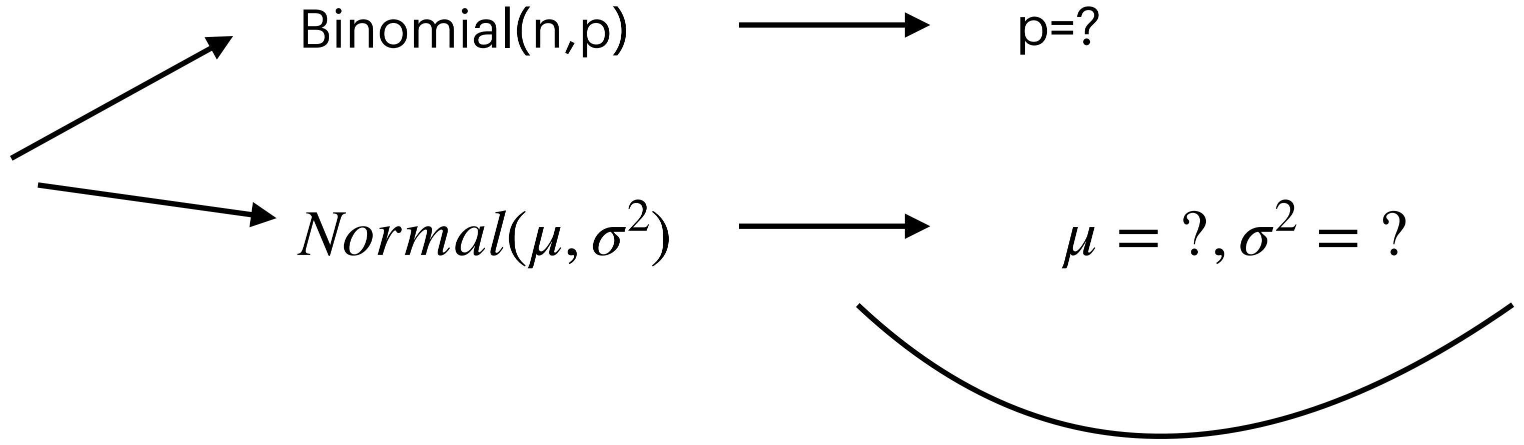
- A home-buyer wants to estiamte the market price for a house before putting out an offer

The houses belong to a population of houses with similar characteristics, with population has some mean μ

Question: $\mu = ?$

Estimation problem

Previous chapter: probability distribution



Binomial(n,p) → p=?

$Normal(\mu, \sigma^2)$ → $\mu = ?, \sigma^2 = ?$

The shape of distribution is determined by parameters

In practice when modeling specific problem,

we need to decide which type of probability distribution is appropriate as a model.

Political poll:
Do you support .... Yes/No

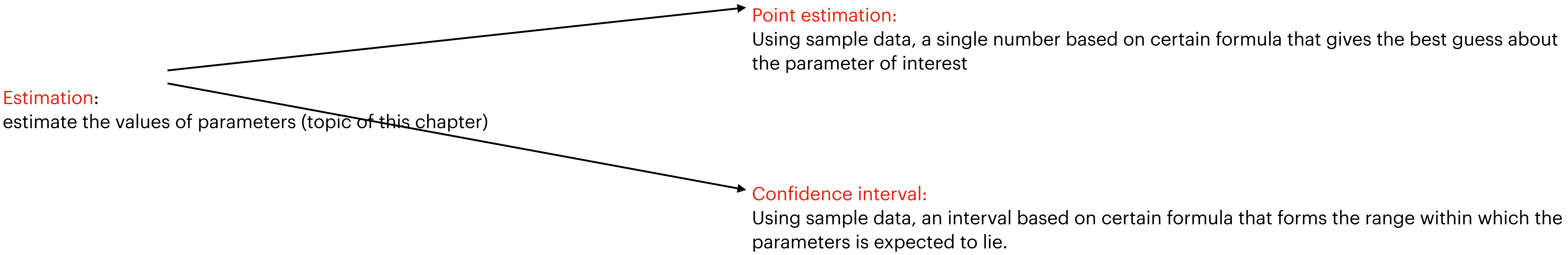
Binomial distribution might be appropriate

Student test scores in a class?

Normal distribution might be appropriate

Once the type of distribution is considered as appropriate,
Still we don’t know the value of the parameters.

We need to “Infer” what the parameters are, i.e. Inference about parameters



Estimation:
estimate the values of parameters (topic of this chapter)

Point estimation:
Using sample data, a single number based on certain formula that gives the best guess about the parameter of interest

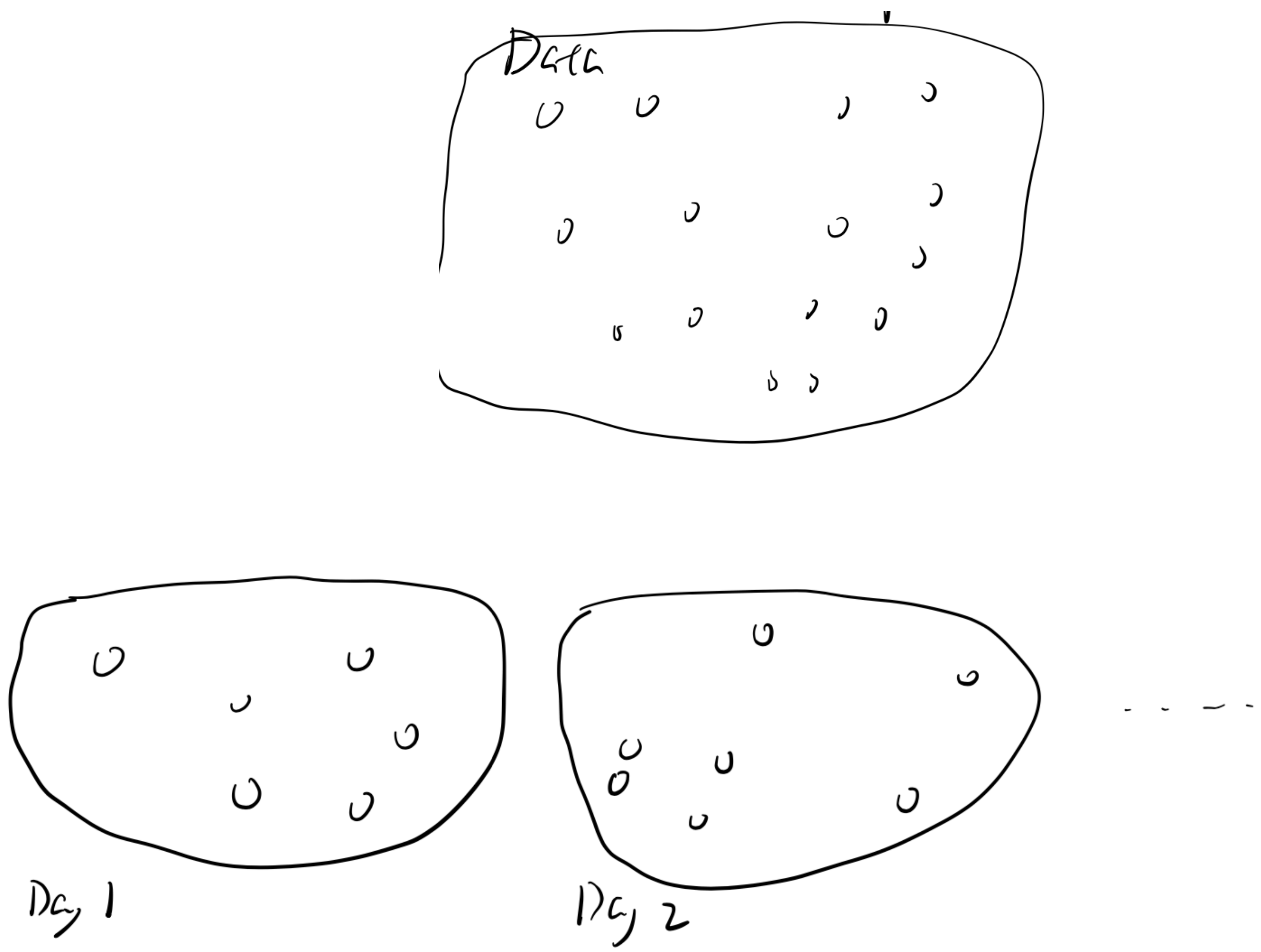
Confidence interval:
Using sample data, an interval based on certain formula that forms the range within which the parameters is expected to lie.

Hypothesis testing:
making decision about whether certain statement is true or false (topic of next chapter)

Sampling Distribution

Once the type of distribution is considered as appropriate,
Still we don’t know the value of the parameters.

Use sampling information, i.e. data, to estimate parameters!



Since you select a random sample, the data or sample will be different each time you re-draw the sample or someone else conducting the same experiment.

➡ So, your estimates will be different, and the difference is the natural variation due to the random sampling process.

When a random sample is drawn from a population, any quantity calculated based on the sample are called a statistic. E.g. sample mean, sample variance, sample proportion...

These statistics change for each different random sample, so statistics themselves are random variables.

Any random variable has probability distribution to describe:

- What values can occur
- How often each value occur

We call the probability distribution for a statistic: the sampling distribution of a statistic

Our goal: sampling distributions for

Sample mean → μ

Sample proportion → p

There are 3 ways to find the sampling distribution of a statistic:

1. Derive the distribution mathematically

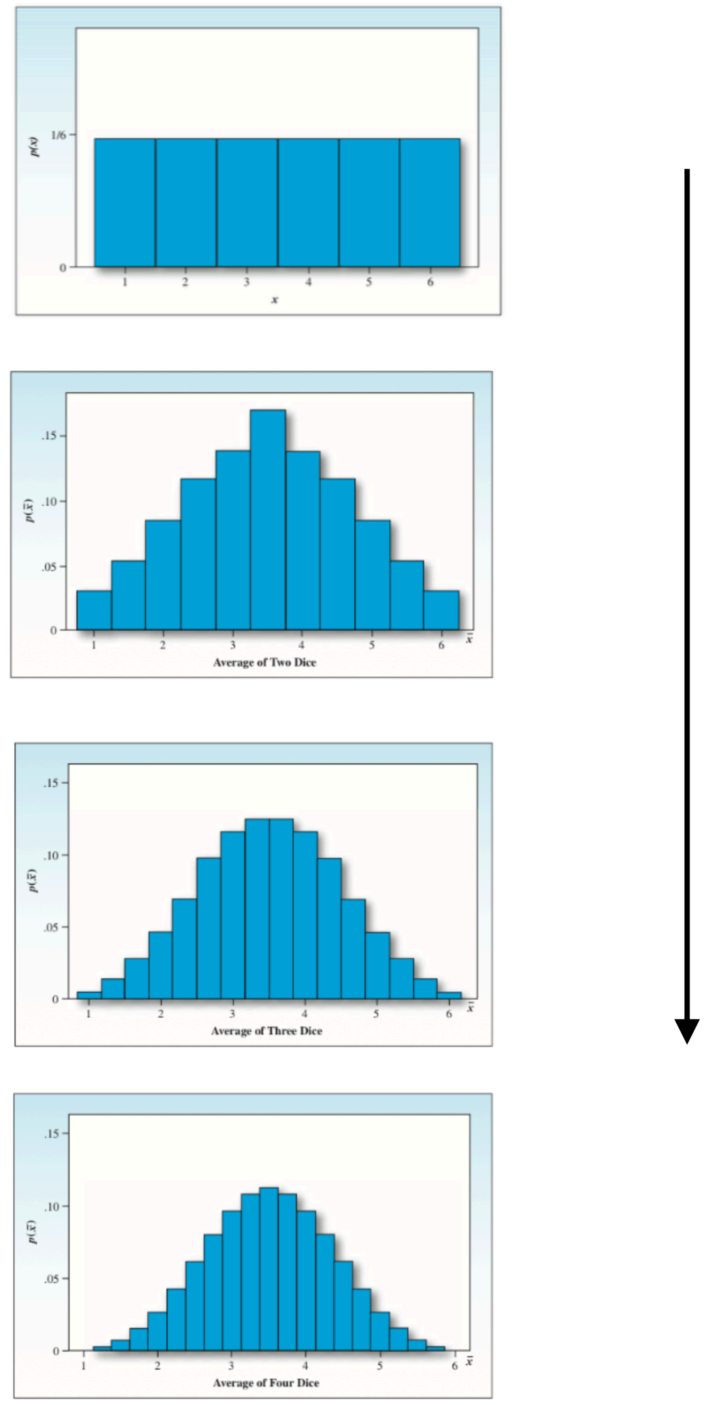
2. Use a simulation to approximate the distribution.

Draw a large number of samples of size n, calculating the value of the statistic for each sample, tabulate the results in a histogram. When the number of repeated sampling is large, the histogram will be very close to the true sampling distribution.

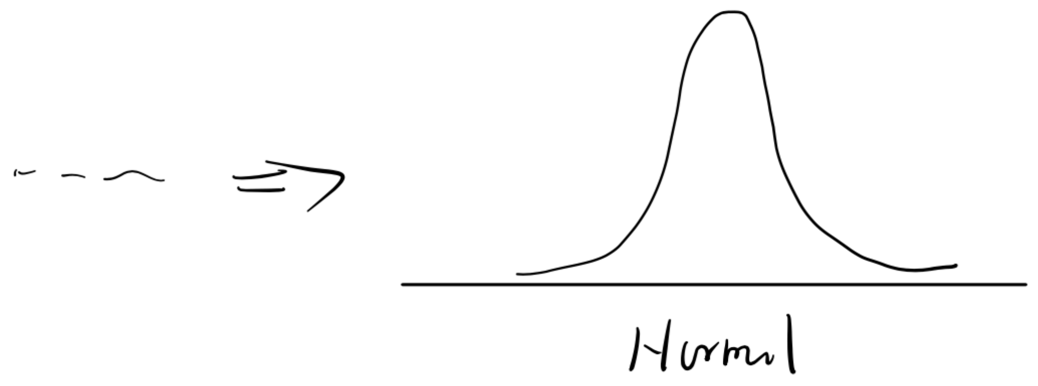
3. Use central limit theorem to derive approximate sampling distribution

Central Limit Theorem (CLT) for Sampling Distribution

What’s the distribution of average number when you toss a fair die many times?



- Symmetric, bell-shaped curve
- Spread of the distribution slowly decreases when we increase n
  - i.e. the distribution becomes thinner, more concentrated around the center



This is true in general:

Average of random samples of measurements drawn from a population tend to have an approximately Normal distribution

Central Limit Theorem

If random sample of n observations are drawn from a population (any distributions, not necessarily Normal), with mean μ and standard deviation σ, then, when n is large (rule of thumb $n \geq 30$):

The sampling distribution of sample mean $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ is approximately normally distributed with mean μ and standard deviation $\frac{\sigma}{\sqrt{n}}$

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

- When the population is actually Normal, then the sampling distribution of sample mean is exactly Normal regardless of how many sample we took
- CLT will be used extensively in the following “inference” problems
- Application of CLT to sampling distribution of the sample mean

The Sampling Distribution of the Sample Mean

If random sample of n observations are drawn from a population (any distributions, not necessarily Normal), with mean μ and standard deviation σ, then, when n is large (rule of thumb $n \geq 30$):

Directly due to the central limit theorem,

The sampling distribution of sample mean $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ is approximately normally distributed with mean μ and standard deviation $\frac{\sigma}{\sqrt{n}}$:

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

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

The standard deviation of a statistic used as an estimator of a population parameter is called the standard error of the estimator, abbreviated as SE

e.g. $SE(\bar{X}) = \frac{\sigma}{n}$

- SE measures: how precise we can estimate the parameter using this statistic, i.e. the precision of the estimator

The duration of Alzheimer's disease from the onset of symptoms until death ranges from 3 to 20 years; the average is 8 years with a standard deviation of 4 years. The administrator of a large medical center randomly selects the medical records of 30 deceased Alzheimer's patients from the medical center's database, and records the average duration. Find the approximate probabilities for these events:

1. The average duration is less than 7 years.
2. The average duration exceeds 7 years.
3. The average duration lies within 1 year of the population mean $\mu = 8$.

To avoid difficulties with the Federal Trade Commission or state and local consumer protection agencies, a beverage bottler must make reasonably certain that 12-ounce bottles actually contain 12 ounces of beverage. To determine whether a bottling machine is working satisfactorily, one bottler randomly samples 10 bottles per hour and measures the amount of beverage in each bottle. The mean $\overline{x}$ of the 10 fill measurements is used to decide whether to readjust the amount of beverage delivered per bottle by the filling machine.

If records show that the amount of fill per bottle is normally distributed, with a standard deviation of .2 ounce, and if the bottling machine is set to produce a mean fill per bottle of 12.1 ounces, what is the approximate probability that the sample mean $\overline{x}$ of the 10 test bottles is less than 12 ounces?

The Sampling Distribution of the Sample Proportion

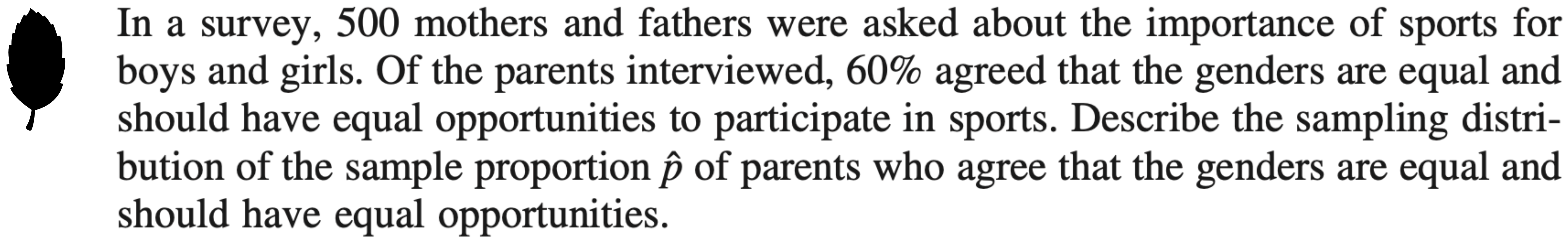
There are many practical situations, such as opinion polls, where we randomly sample n people to estimate the proportion p of people in the population who have a specific characteristic.

X= total number of sampled individual who have this characteristic

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

Then directly due to the central limit theorem, then the sampling distribution of sample proportion can be approximated by a normal distribution:

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

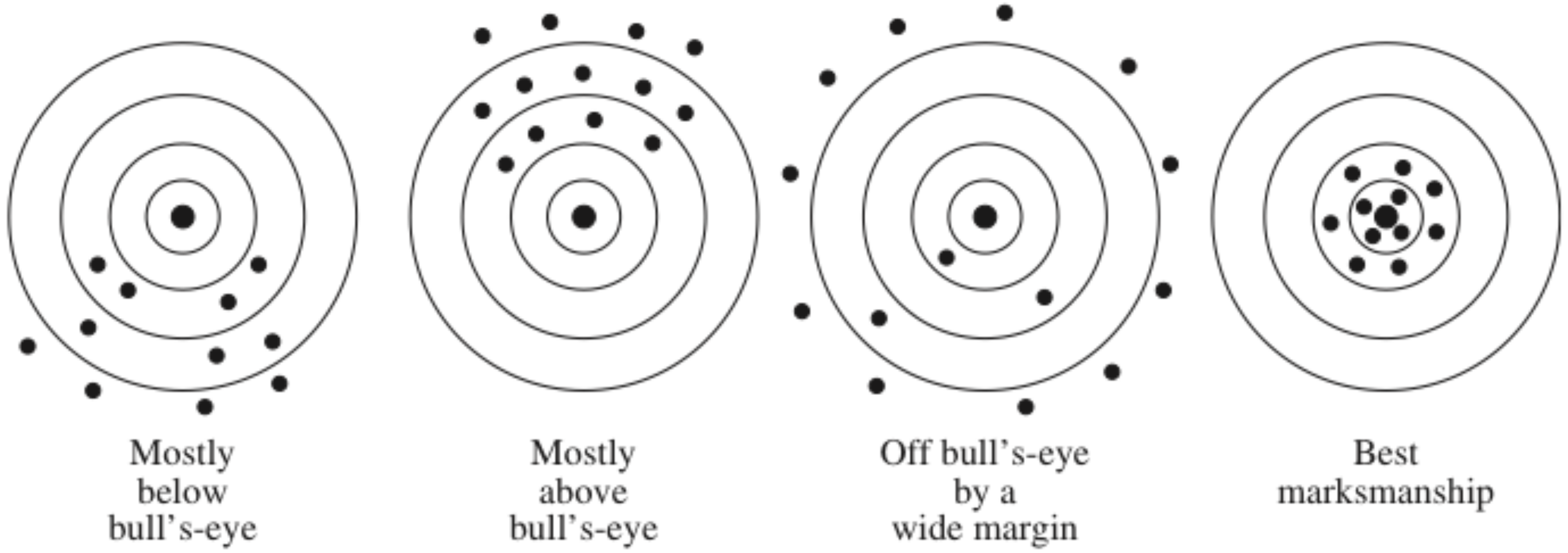


In a survey, 500 mothers and fathers were asked about the importance of sports for boys and girls. Of the parents interviewed, 60% agreed that the genders are equal and should have equal opportunities to participate in sports. Describe the sampling distribution of the sample proportion $\hat{p}$ of parents who agree that the genders are equal and should have equal opportunities.

If the actual p=0.55. What is the probability of observing a sample proportion as large as or even larger than the sample proportion 0.6?

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

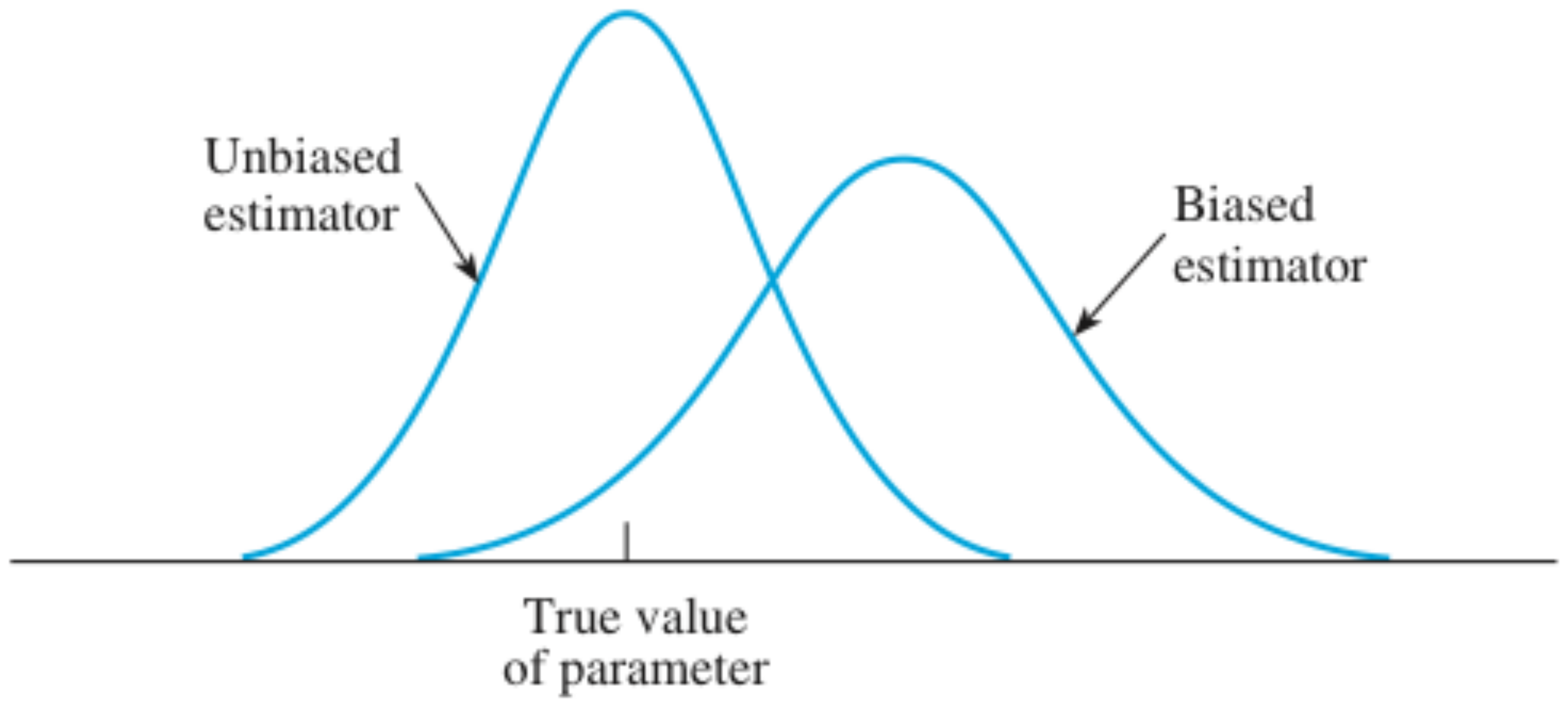
Which one has the best shot?

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.

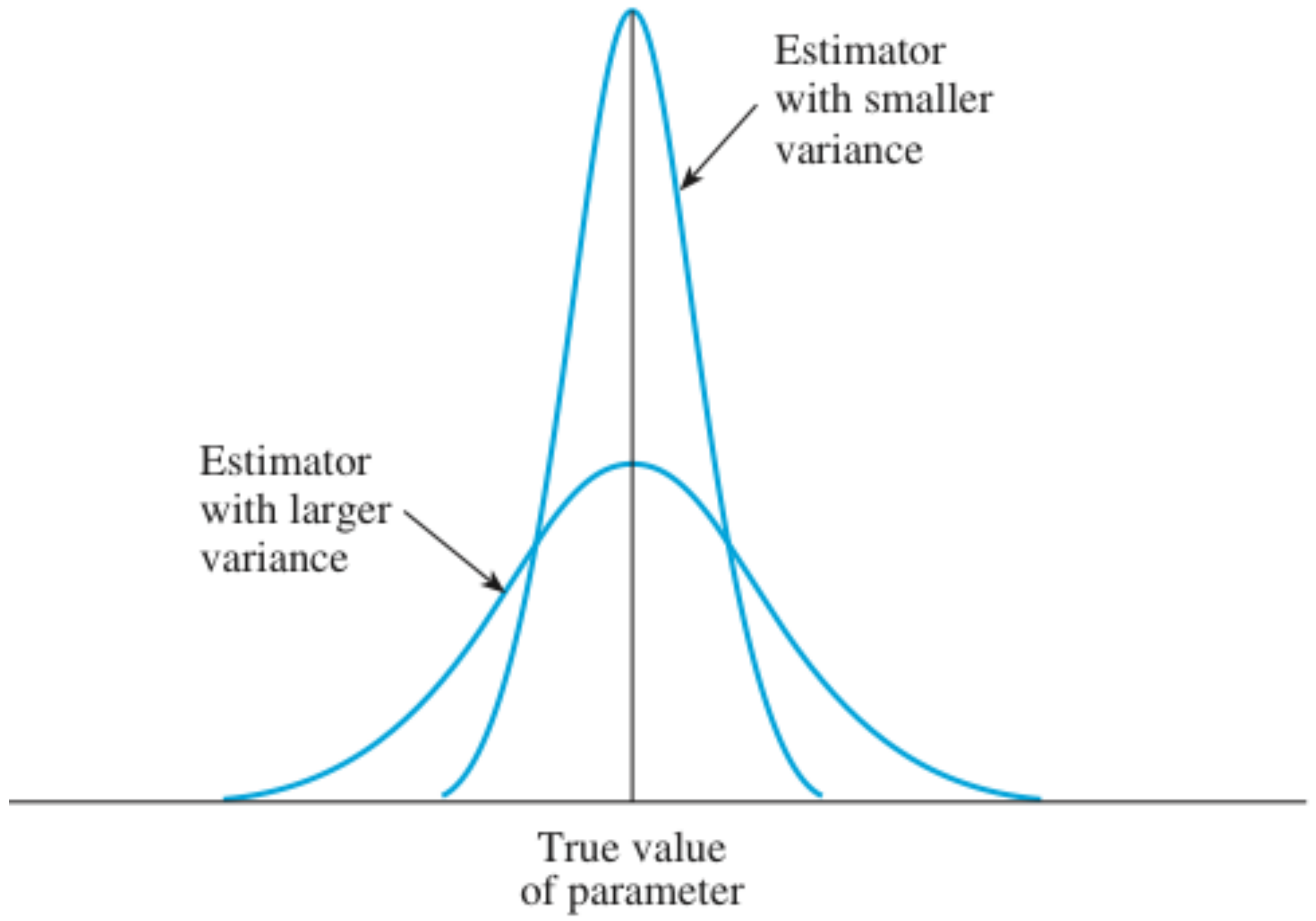
➡ Same principle applies to point estimation

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


Estimator with smaller variance
Estimator with larger variance
True value of parameter

How to estimate a population mean?

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

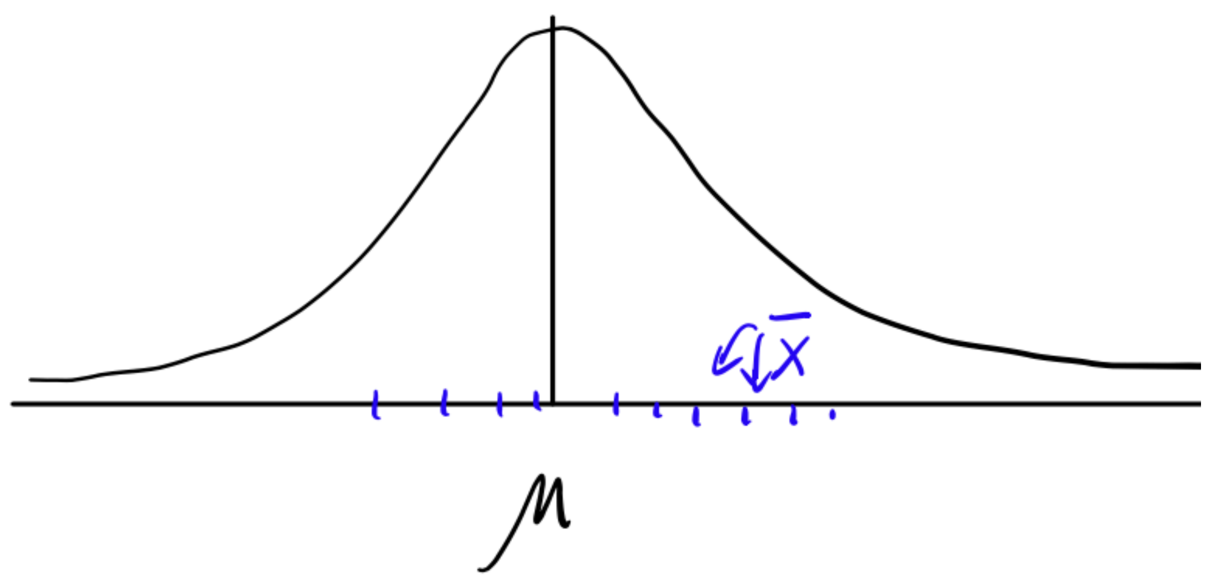
Then, the point estimator for the mean μ is the sample mean $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$

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:

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

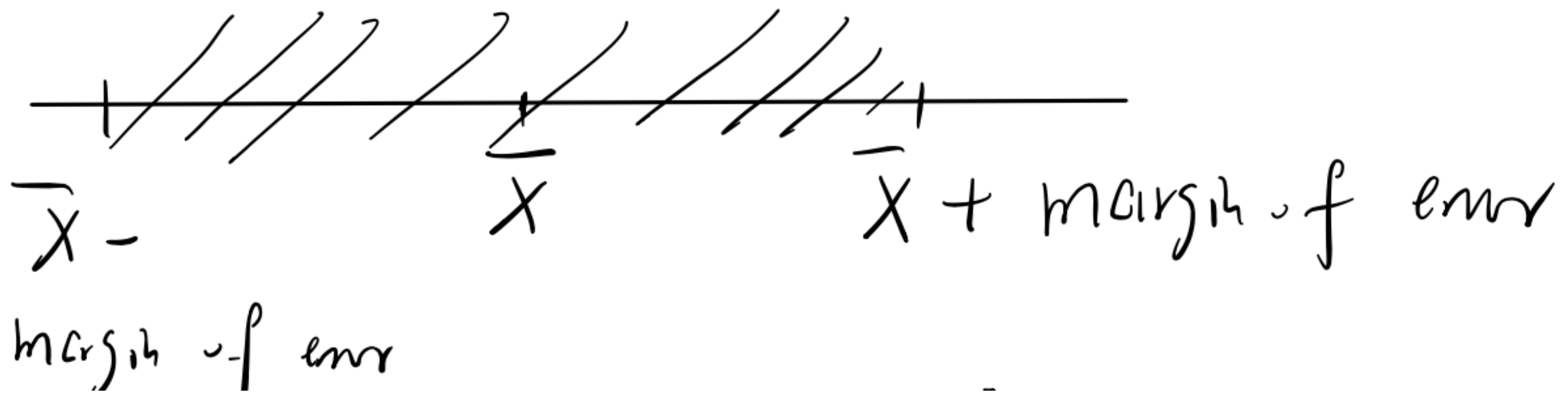


Due to the fact of Normal distribution or the 68-95-99.7% rule:

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

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.

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.

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.

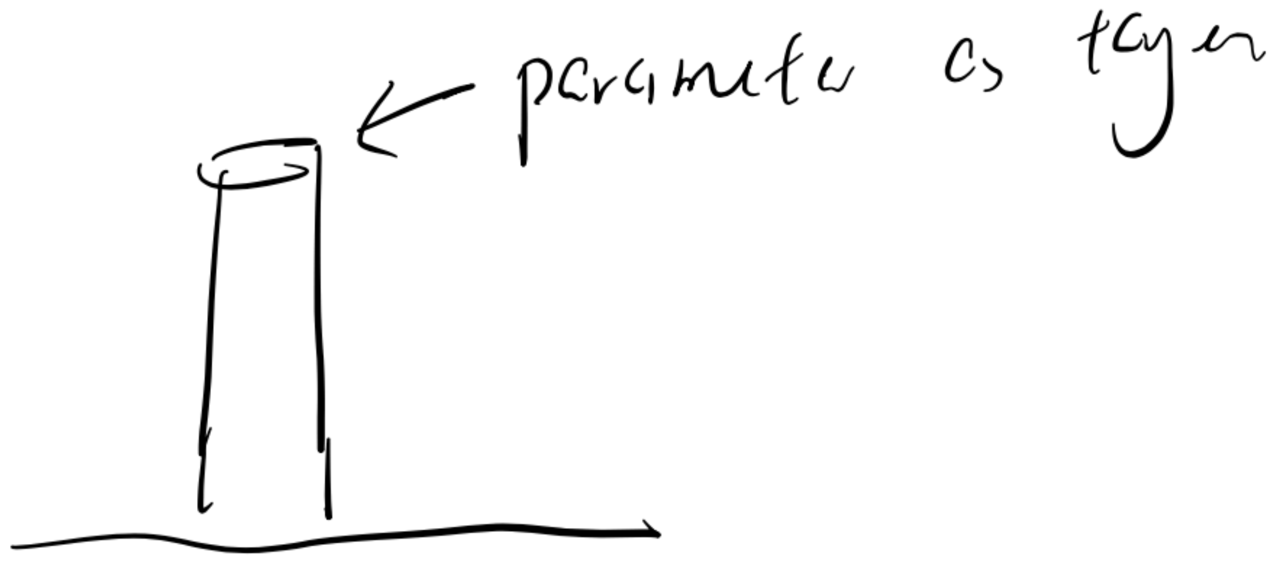
If the data provide more information, then we’re more certain about our point estimate, and the margin of error will be smaller; otherwise, the margin of error will be larger, reflecting more uncertainty.

➡ 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

Miss vs Catch :

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 %)

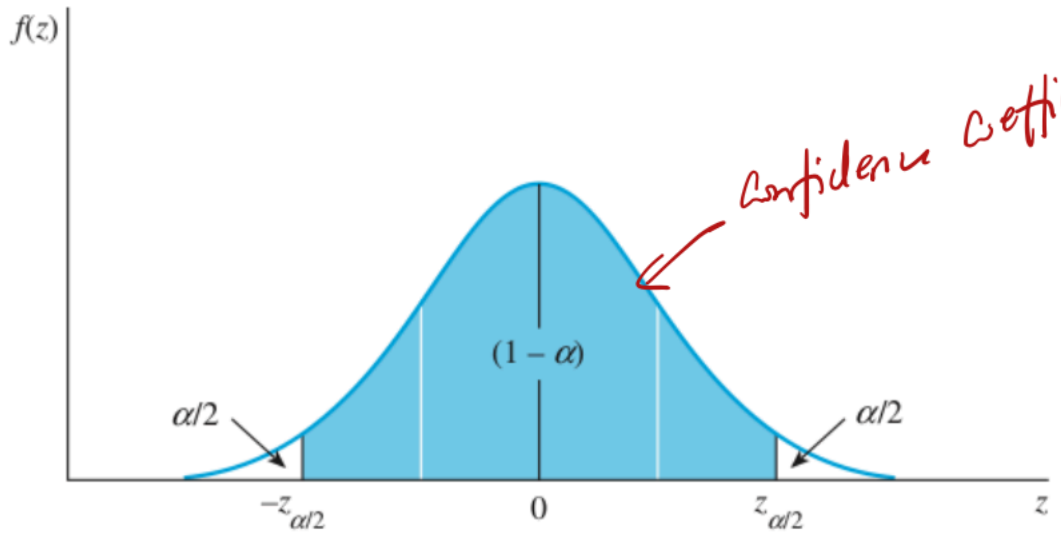
- If too small: often miss the target
- Usually: 0.9, 0.95, 0.99

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

$$(\text{point estimator}) \pm \underbrace{\underbrace{Z_{\alpha/2}}_{\text{the value of } z \text{ that has tail area } \frac{\alpha}{2} \text{ to its right}} \times (\text{standard error of the estimator})}_{= \text{ margin of error !}}$$

$$= [\ \underset{\uparrow}{LCL}\ ,\ \underset{\uparrow}{UCL}\ ]$$

lower confidence limit   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

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

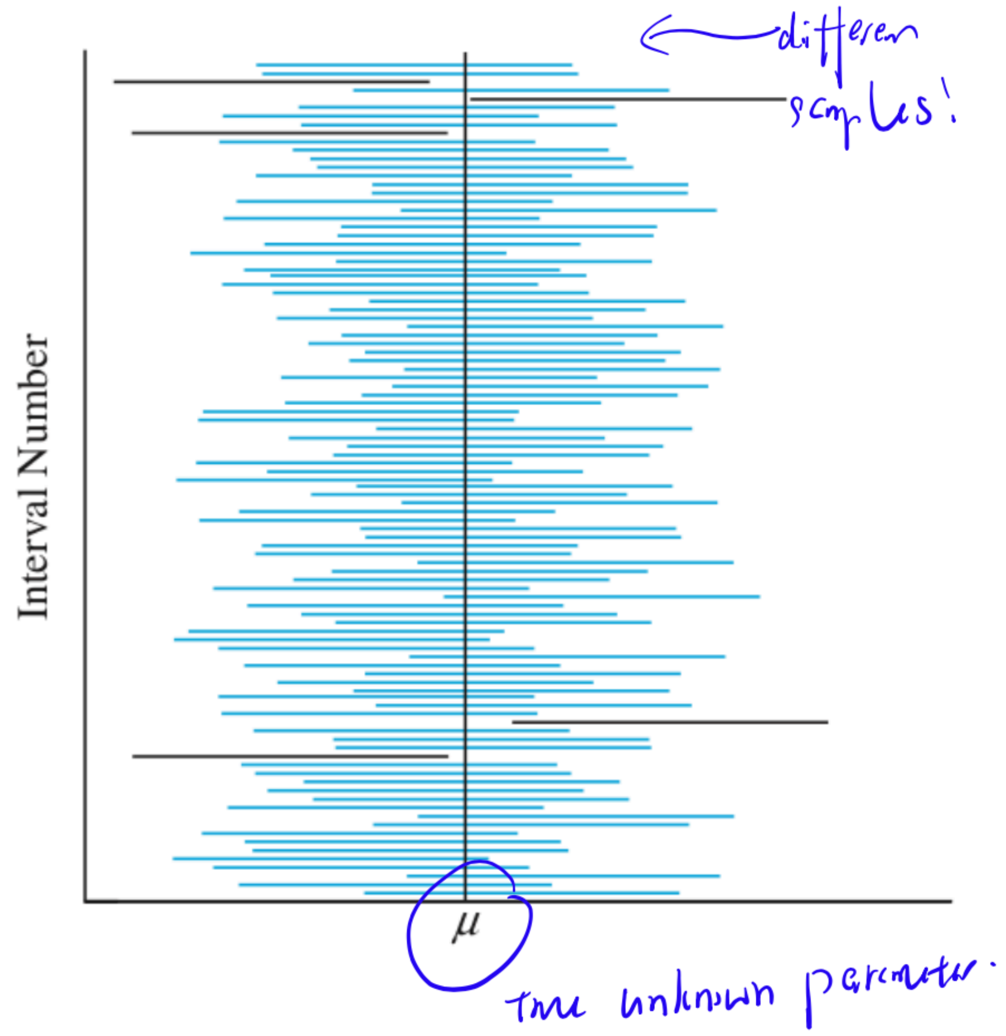
If σ is unknown, replace with sample standard deviation $S = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2}$

$(1-\alpha)100\,\%$ 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.

universe 1
universe 2
universe 3
universe 4
. . . . . .

A dietician selected a random sample of $n = 50$ male adults and found that their average daily intake of dairy products was $\overline{x} = 756$ 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.

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?

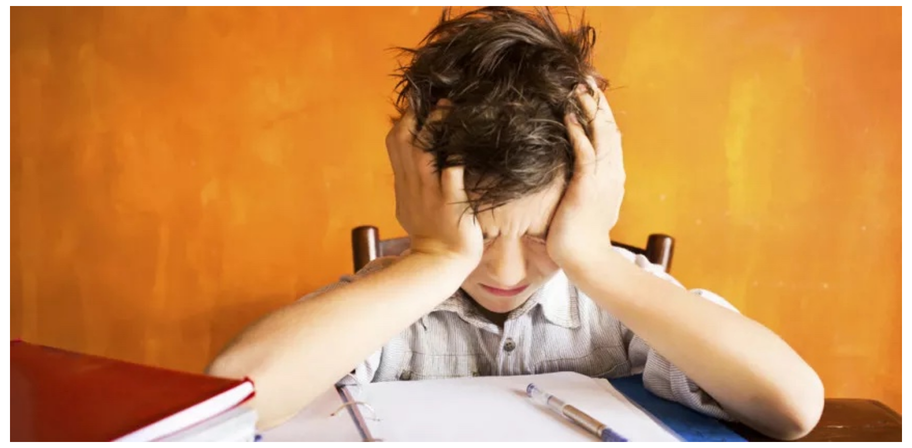
biased sample

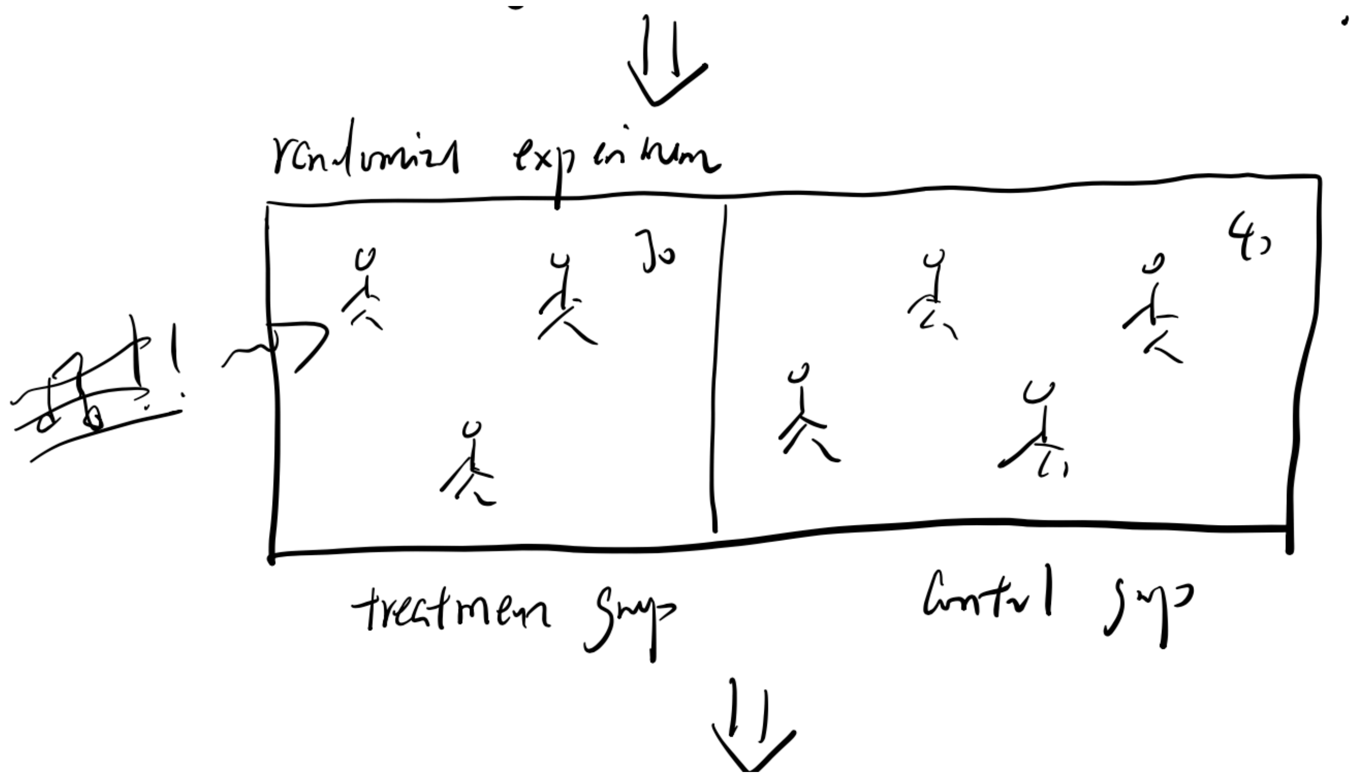
⇒ 90% confidence

X

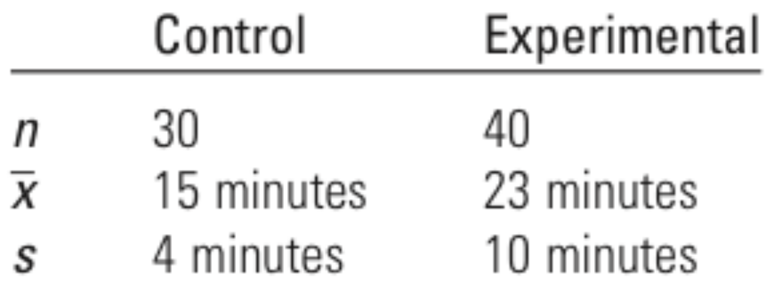
Noise and Stress

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



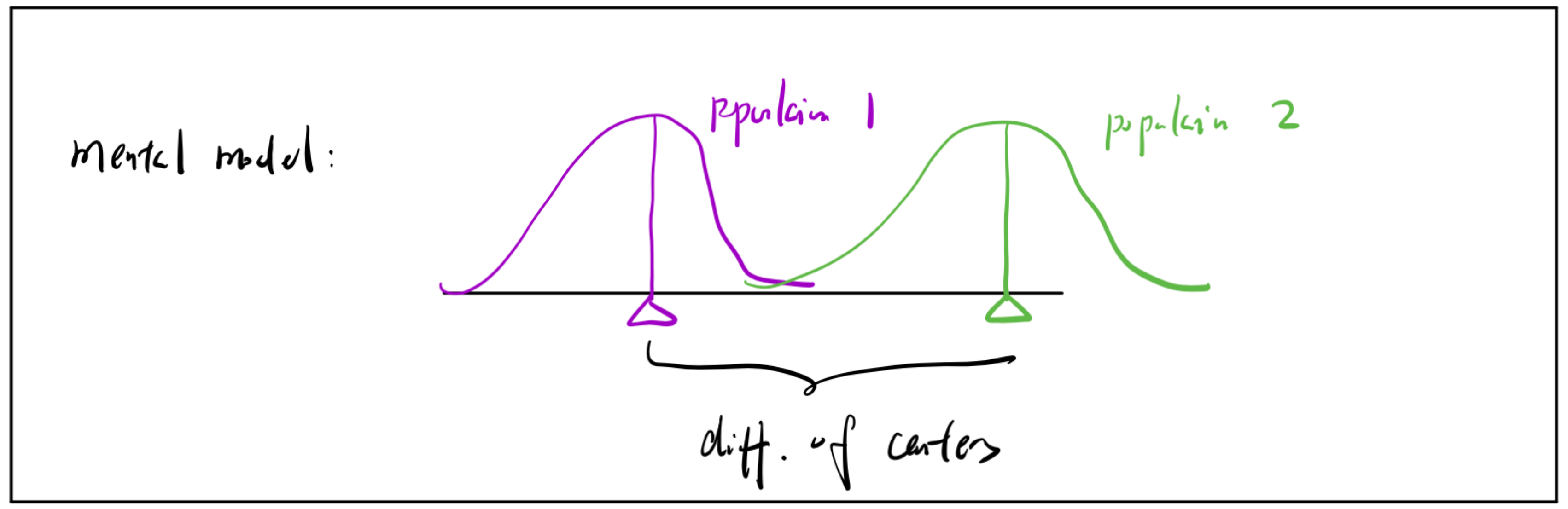


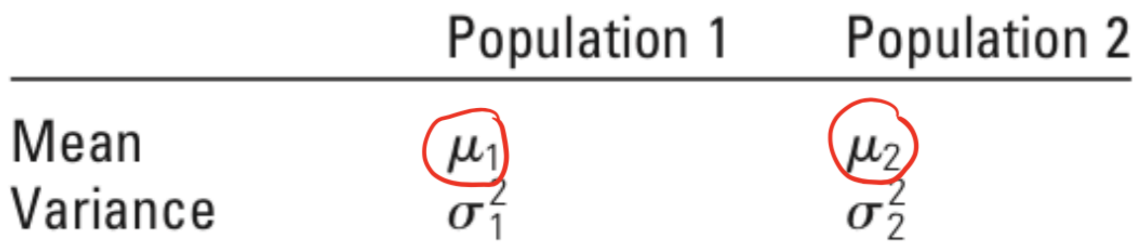
Record “time to finish the task”



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

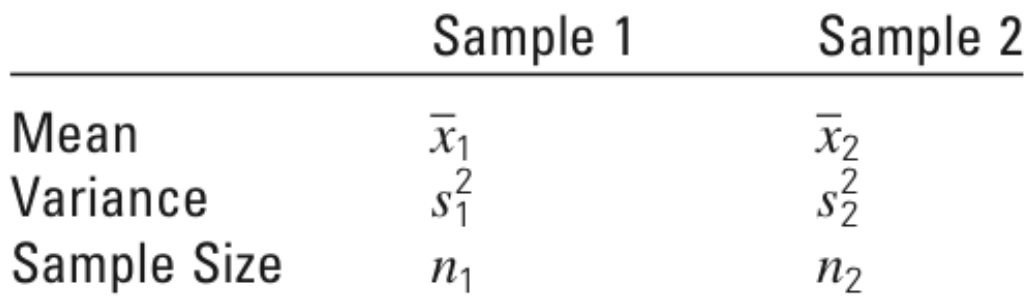
Many questions are comparison between two populations, specifically, looking at the difference between two population means. E.g. is the new vaccine more effective than the old one? ....





	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

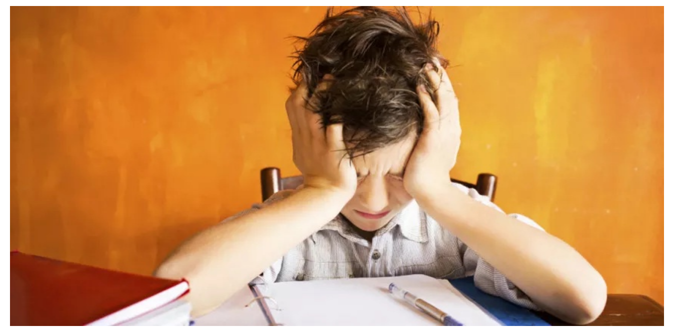
3. 95% margin of error: $1.96\sqrt{\frac{\sigma_1^2}{n_1} + \frac{{\sigma_2}^2}{n_2}}$

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

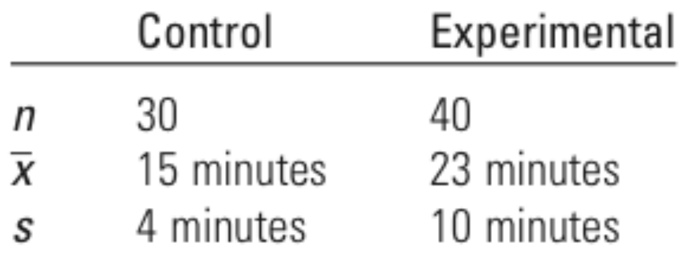
$$\bar{X}_1 - \bar{X}_2 \pm z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{{\sigma_2}^2}{n_2}}$$

Noise and Stress

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



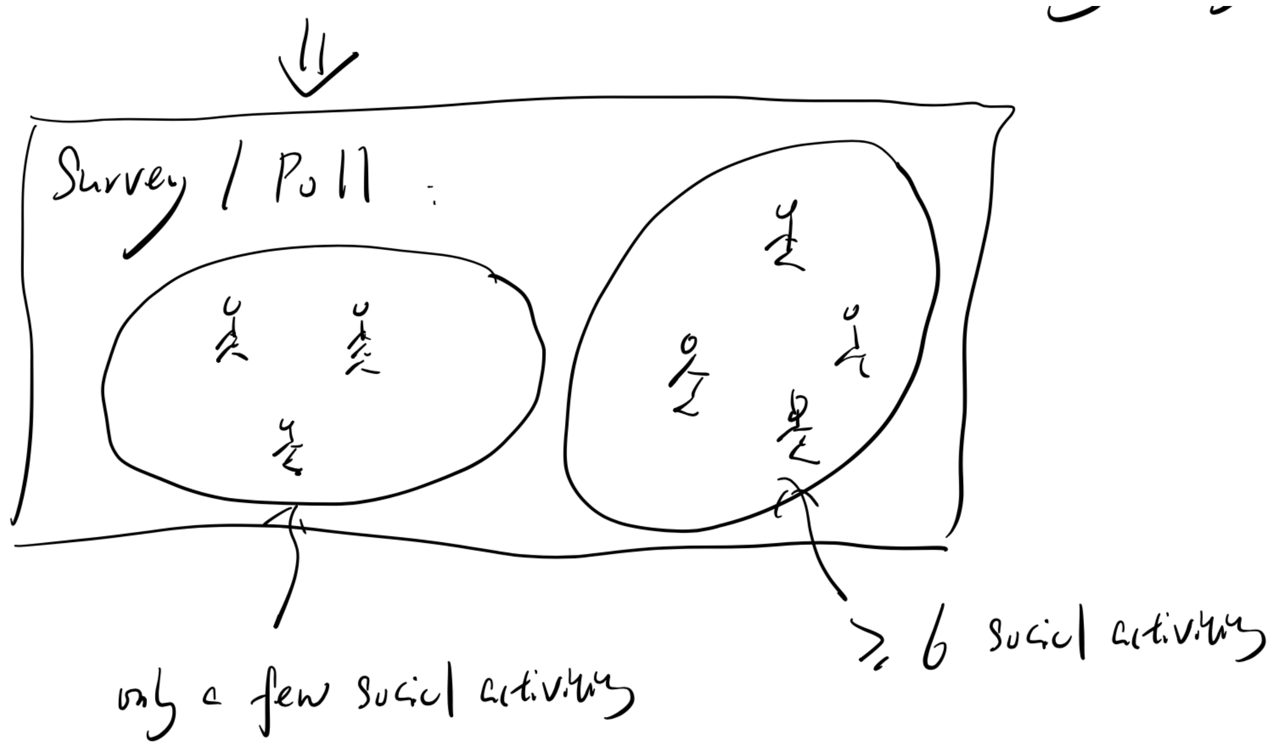
Record “time to finish the task”

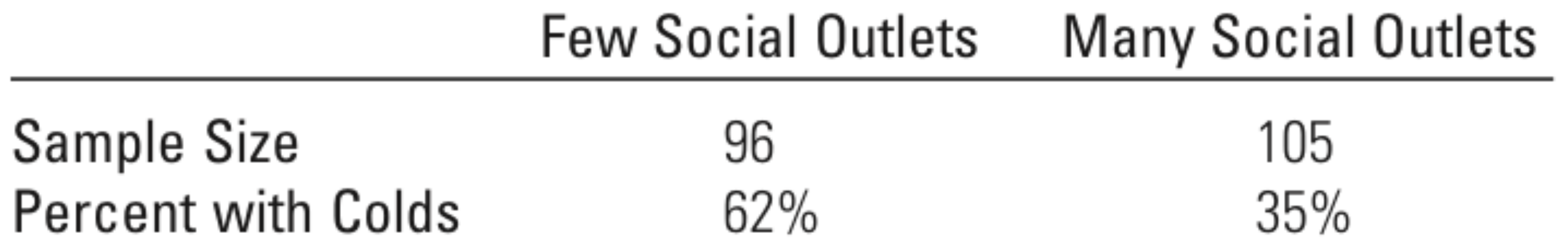


	Control	Experimental
n	30	40
$\bar{x}$	15 minutes	23 minutes
s	4 minutes	10 minutes

Social activities vs Getting colds

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

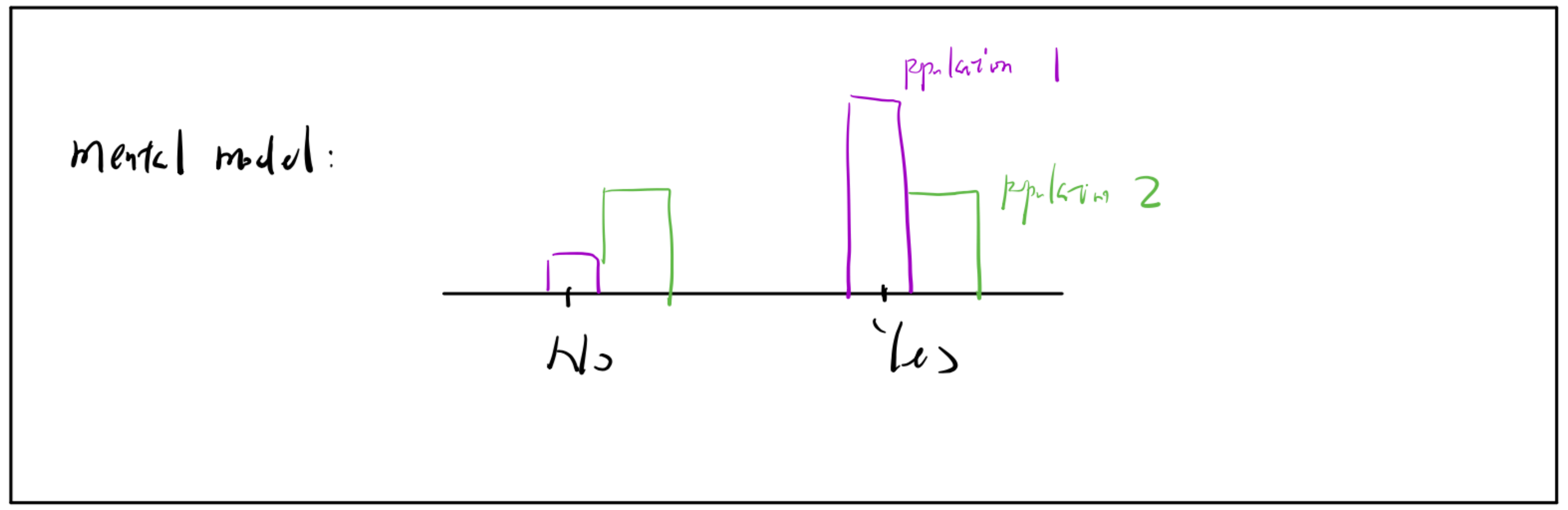




	Few Social Outlets	Many Social Outlets
Sample Size	96	105
Percent with Colds	62%	35%

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

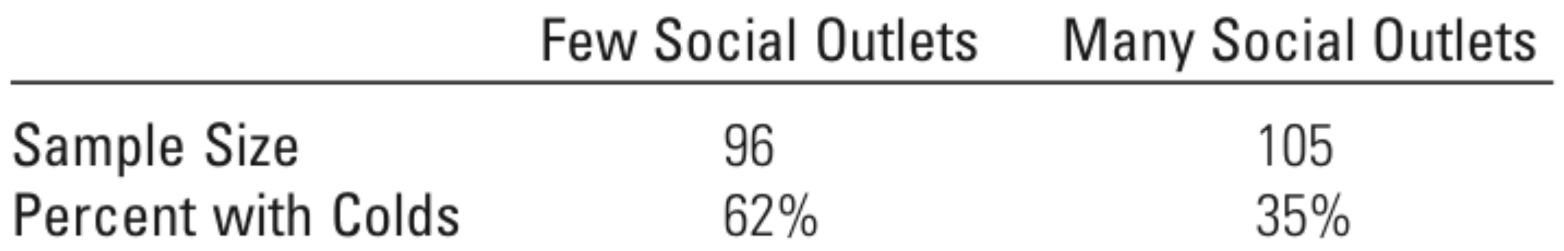
3. 95% margin of error: $\pm 1.96\sqrt{\frac{\hat{p}_1\left(1-\hat{p}_1\right)}{n_1} + \frac{\hat{p}_2\left(1-\hat{p}_2\right)}{n_2}}$

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%