Chapter 1

Planning a Statistical Study

О достоверности опросов можно судить с поправкой на то, что люди лгут через слово.

Unknown author

Obtaining Data

No study is useful without data. Researchers may be interested in many different questions. What proportion of citizens will support the political party on the elections? What is the real effect of vitamin C on human health? What is the effect of help rooms for students on their exam scores? How to answer them? For this we need to gather and analyze the data on the corresponding topics. Or the researchers may already have some suppositions about reality. Again any hypothesis should be verified by the real data. So, all the studies are based on information obtained from the reality. It is crucial to make sure that your data is reliable. It is not only analyzing, but also gathering of data that should be done in a proper way. In this chapter we will discuss how to correctly obtain the data for statistical studies.

The group of all units we are interested in in our study is called population. It is a general term. For instance, if we were to study nonsmokers in a country, we would refer to them as a study population, and will not consider smokers.

Observational Study versus Experiment

There are two different approaches to statistical studies: observational study and experiment. They differ essentially in the way they are organized. In an experiment we *impose a difference* in conditions and then measure how it affects the result, while in observational study we observe subjects/units which are already *in different conditions* and take notice of how this difference is related to the result.

For example, we might want to study the dependence of body shape on sport exercises. An observational study would be to *find* 100 people, ask them about whether they practise exercises, measure their body condition, and then evaluate how it is related to exercising. An experiment would be to *take* 100 people, 50 of which are assigned to practice exercises, and the other 50 – not to practise. Then,

after 1 year we measure the change in shape of bodies in each group and conclude on how the shape depends on practicing exercises.

Which approach to choose? The answer depends on the particular research questions, as well as on the budget and ethical constraints. Generally the experiment is better whenever possible.

Observational Studies

Sources of data for observational studies include census and sample surveys. We will point out about the first and then specifically analyze the latter.

Census

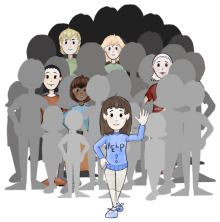
Census means gathering data about the *whole* population. Every 10 years the Federal State Statistics Service of Russia attempts to gather information about every citizen of the country. A massive amount of data is obtained. Basically, even the census may not include every citizen. After completion there always exist households, which were not reached, as well as missed homeless people.

Usually, census is much too expensive and time consuming. It is appropriate only if it has a reason or if the population involved in a study is rather small and every member is reachable. For instance, if you are interested in only one academic group it is easy to ask every student in this group.

Sample Surveys

Sample survey aims to obtain information about the whole population by taking and studying a part of it, called a sample. Number of elements in the sample is called the sample size and is usually denoted by n. In Chapter 5 we made a differentiation between sample and population for the first time. Now let us discuss it in more detail.

Why do we *need* sampling? Imagine that a manufacturer wants to bring a new model of a Smartphone to a market. He wants to estimate the potential current demand which will determine his profit by conducting a market research.



Manufacturer is interested in the population of all potential buyers. Obviously, it is impossible to contact *every* potential customer. Moreover, imagine how long it would take to make the population census of the whole country or even of the several countries! In that time there would be released two new models of that Smartphone and the current model would be no longer valid. So this market research would be already useless.

So a rather small subset of population members is contacted, namely a sample. Then the conclusion about the population would be made as a *generalized* conclusion about the sample.

Sampling Error

All conclusions based on sample surveys contain some level of *uncertainty*. It is a fact

that has to be accepted. It is natural and arises because characteristics of a sample do not equal exactly the same characteristics of population. Fortunately, this is what Statistics is all about, measuring and taking into account the uncertainty.

No matter how well-designed and well-conducted a sample survey is, it still relies on characteristics of a sample, not of the whole population. For example, three different researchers might independently report the average monthly expenses of an ICEF student are 22500, 18700 and 21500 roubles, correspondingly. All three numbers estimate the real average expenses, which could be some other forth number, not equal exactly to any of these estimates. The effect of such difference between true and estimated parameters is called **sampling error**. It is naturally present and the term *error* doesn't mean something is wrong. Statistical theory allows describing the size and behavior of sampling error. Taking that into account we can still make valid statements about the population.

So, sampling error is always present. It can only be decreased by increasing the *sample size*, but not reduced to zero.

Why sample has to be random?
Since only a part of the population is taken (sample) and conclusions will be based only on it, it is necessary that the sample is as *representative* of the population as possible. The word 'representative' means that the sample reflects all the basic properties of the population. For example, representative sample from the population of the Netherlands should include many people who are tall, have blond hair and light-colored eyes.

Representativeness of a sample is achieved by *random* choice of objects from the population. Thus, for the Netherlands population which contains many tall people, the chance of randomly choosing a person, who is tall, is also high. So, a random sample is expected to contain many tall people as well. Randomness ensures representativeness, that's why all probability sampling methods rely on random sampling.

Types of Sampling Methods

1. Simple random sampling (SRS)
Simple random sampling is the most popular and basic procedure. By definition SRS ensures that every sample of size n has an equal chance of being selected. This is achieved by choosing each individual entirely by chance. In SRS each element has an equal chance of being selected *and* each group of size n has an equal chance to become a sample.

You can think of SRS as of "the Hat Game". Each population member is given a number. Numbers are written on papers, put into a hat, mixed and then the required number of papers n is pulled out with closed eyes. Individuals with corresponding numbers are chosen into a sample. This procedure provides a simple random sample of size n.

Practically, to get SRS random number table or random number generator is used.

84177	06757	17613	15582	51506	81435	41050	92031	06449	05059
59884	31180	53115	84469	94868	57967	05811	84514	75011	13006
63395	55041	15866	06589	13119	71020	85940	91932	06488	74987
54355	52704	90359	02649	47496	71567	94268	08844	26294	64759
08989	57024	97284	00637	89283	03514	59195	07635	03309	72605
29357	23737	67881	03668	33876	35841	52869	23114	15864	38942

Figure 1.1: random number table

How to use random number table

Imagine that you need to randomly select a group of 12 students out of a class of total 50 students. Below are the steps for the full procedure.

Assign all students from 01 to 50.

Read off two-digits numbers from a random number table, ignoring repeats and out of order, until you get 12.

For example, looking at a given table the 2-digit integers are 84, 17, 70, 67, 57, 17, 61, 31, 55, 82... The first number that fits is 17. The next is 31. So, 17 appears again, but it was already chosen into the sampe and cannot be taken for the second time.

The students corresponding to those 12 integers will be chosen for the sample.

More up to date practice is the random number generator. In fact, a table is just a realization of a generated order of random numbers.

The main advantage of SRS is its simplicity. Such samples are easy to work with. Therefore most of the statistical procedures (e.g. statistical tests, explained in Chapter 11) are based on the assumption that the sample taken is SRS.

However in some cases SRS is not appropriate. For such situations there exist alternative sampling techniques described below.

2. Stratified random sampling

This type of sampling exploits grouping of population members within the population. It is needed when population members differ in some characteristic, which affects study results. Such population is called heterogeneous. For example, while studying behavior of students in international university, it is necessary to make sure that representatives of students from all the countries are present in the sample.

We first stratify, that is divide, the entire population by the characteristic of interest (age, gender, preference) in strata. Note that individuals in each stratum will have the same level of the characteristic of interest. This will make strata homogeneous. Then we take simple random samples from each stratum.

Simple random sampling is the most appropriate when the entire population is homogeneous. An SRS can miss "unpopular" individuals – minorities, although characteristics of these individuals may affect the result of the study. That is why when the population is heterogeneous SRS is not appropriate, and stratification should be used.

Proportional sampling is a special case of stratified sampling when the size of a random sample from each stratum is proportional to the size of this stratum in the entire population.

3. Systematic random sampling

This type of sampling involves picking every *n-th* person starting from some random point. For example, to form a sample of night club visitors we can pick every 5-th person getting out of the club, if the population of study is night club visitors.

It is easy to implement, although it should be used carefully. The problem is that some data might have periodical style. For example, if you are interested in total sales per day during a month, sales tend to be similar on all Mondays, or all Sundays. If you take every 7-th day, this will result in a non-representative sample.

4. Cluster sampling

Sometimes the homogenous population is divided into isolated groups, but not with respect to some characteristic of interest. For example, the division may be based on geographical or language areas, like districts of a city or regions of a country. In such cases taking SRS will result in subjects being spread over a large area, and contacting each of them would be too costly. So, an alternative sampling procedure may be used, called cluster sampling. This approach is attractive when the entire population can be conveniently subdivided into isolated groups called clusters. Then, a simple random sample *of clusters* is taken from the population and all the elements in each chosen cluster are included in the sample. If the sampling is done in two stages, then at the second stage a SRS is taken from each chosen cluster.

Cluster sampling technique may be more practical and/or cheaper than simple random sampling. For example, we are interested in the use of pesticides by farmers in some region. We check the region and take a simple random sample of 15 villages. After that we include all the farmers from each chosen village. In this way, we will not need to run around the region to contact several farmers

Please, do not mix it up with *stratified* sampling! Clusters do *not* differ with respect to a characteristic of interest. It is just more convenient to divide a homogeneous population into clusters. Contrary, strata are parts of heterogeneous population and vary with respect to an important characteristic. The main aim of cluster sampling is to reduce costs. This contrasts with stratified sampling where the main objective is to increase precision.

5. Multistage sampling

This type of sampling refers to a procedure involving two or more stages, each of which could use any of different sampling techniques presented above.

All the methods discussed above are the *probability* sampling methods. That means every population member has a positive and known probability to be chosen. These methods, except for the systematic sampling, are possible only if researchers have sampling frame, which is the whole list of population members. This is the main feature of a random sample. SRS is an extreme case of random sampling when all members have an equal chance of being selected. Distinguish between 'sampling' and

'sample'. Sampling is a procedure of choosing objects, and sample is the resulting set of objects chosen.

Visually sampling methods can be represented in the following way:

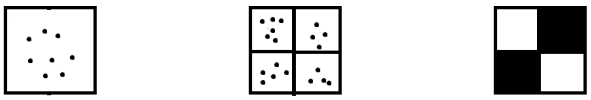


Figure 1.2: sampling methods: simple random, stratified, cluster

"Bad" sampling methods

However, there exist other sampling techniques used in practice that suffer from methodological defects. Their drawbacks are the following:

- some population members have zero chance of being selected
- for population members that can be selected it is impossible to determine the probability of being selected
- sampling error cannot be quantified

Usually those techniques are used when the sampling frame is unavailable. Consider the following examples:

- **Convenience Sampling** takes units that are easy to reach. For instance, if we want to investigate the opinions about the quality of food at the university canteen and we ask students currently present there. In this way we ignore the opinion of those who are absent. It is very probably they dislike the food and this is why did not come.
- **Quota sampling** attempts to obtain representative sample by specifying quotas on members in sample with certain characteristics. The distribution of such characteristics in population is required in order to follow it in the sample. For example, we want to study the opinion of teenagers on a certain movie. We know that in general preferences of boys and girls are different and thus we want to have both represented in the sample. If we want a sample of size 20, we need 10 teenage girls and 10 teenage boys. Interviewers will survey all subjects they can reach until the required level of quota is achieved, no matter how they choose those subjects. This method is widely used in opinion polling or market research. Obviously, the concept of randomness is violated.

Note that none of the alternative procedures will result in a simple random sample (SRS), because every possible sample of a size n does not have an equal chance of being selected.

Bias

A good sample survey is conducted in two steps:

1. Take a right sample.
2. Ask right questions.

If any of these steps is done wrong, the results of the survey will be misleading. Then, we say the survey has a bias.

Errors caused by bias represent another, non-sampling type of errors. They do not depend on how large the sample size is. If data were obtained in a wrong way, bias could happen even if a study uses the whole census. The whole population may be contacted but bias will still arise due to mistakes in the procedure.

There could be several sources of bias in one survey. There are two classes of bias: selection bias and response bias.

Selection bias: "wrong sample"

As we have already said a good sample is representative. If sample does not accurately represent the population, then the researcher has chosen a wrong sample. The bias that results from an unrepresentative sample is called the **selection bias**.

Literary Digest. This is a classic example of study with a selection bias. In 1936 a popular weekly magazine Literary Digest conducted a poll of voters and predicted that Alfred Landon would beat Franklin Roosevelt in the upcoming presidential elections. They created a mailing list of 10 million people and sent them a mock ballot via mail asking to choose the preferred candidate and send the ballot back. As a result, they received 2 million responses and concluded that Landon would win. However, this mailing list was drawn from telephone directories, car registration lists, lists of magazine subscribers etc. There were two substantial problems with this survey. First, in 1936, people who owned cars and telephones tended to be richer and thus the resulting convenience sample ignored the opinion of poor people. Second, low response rate introduced additional bias to the results.

- **Undercoverage bias** occurs when some members of the population are inadequately represented in the sample. It is well illustrated by the Literary Digest example. The survey sample suffered from undercoverage of low-income voters, who tended to be Democrats. Thus, poor people were by design excluded from the sample.

  How did this happen? The survey relied on a *convenience sample*, drawn from telephone directories and car registration lists. In 1936, people who owned cars and telephones tended to be more rich. Convenience samples often lead to undercoverage bias.

- **Nonresponse bias.** happens because the characteristic of interest differs for respondents and non-respondents. Sometimes, some part of individuals chosen for a survey is unwilling or unable to answer. And it could be that those who did answer differ in a meaningful way from those who did not. Coming back to our example, the respondents in the Literary Digest survey tended to

support Landon. On the other hand, the non-respondents mostly supported Roosevelt as it was learnt after. Since only 25 % of the chosen voters actually completed the mail-in survey, the portion of Landon supporters was biased upwards. Namely, the results overestimated support for Alfred Landon.

Note the difference between these two types of biases. Undercoverage bias happens because we *didn't contact* some important part of population. Contrary, non-response bias happens because some contacted part of population *didn't answer*. This is a common problem of mail surveys. Response rate in such surveys is often too low which makes non-response bias present.

- **Voluntary response bias** also happens since respondents differ from non-respondents. An example could be a TV or radio show surveys which ask questions online. Often the topics are highly controversial, like prohibition of abortion, immigration or gun control. The members of such surveys are self-selected. Usually they have a reason of strong will to participate in it, otherwise they will not call. This is called strong opinion. So usually the result will over-represent individuals with strong opinions on the topic concerned.

Response bias: "wrong questions"
Another problem is the way the information is gathered from the individuals who are already in the sample. The way of obtaining information (the measurement process) includes the environment in which the survey is conducted, the way questions are asked, and the state of mind of the survey respondent. A poor measurement process leads to response bias.

- **Non-neutral questions**. The question that respondent is asked may be made in such way to unduly favor one response over another. For example, a satisfaction survey may ask the respondent to indicate whether she is satisfied, dissatisfied, or very dissatisfied. By giving the respondent one response option to express satisfaction and two response options to express dissatisfaction, this survey question is biased towards getting a dissatisfied response.

- **Non-anonymous surveys.** Most people like to present themselves in a favorable light, so they will be reluctant to admit unpleasant attitudes or illegal activities. Instead, their responses may be biased towards what they believe is *socially desirable*. For instance, there may be a question regarding an attitude towards racism or homosexuality, with the section to write your name as well.

  A nice example of bias occurred in non-anonymous survey is given by the following case. 221 out of 334 people who named themselves vegetarians in one survey happened to admit eating meat or fish during the last month in the other survey. `https://www.ncbi.nlm.nih.gov/pubmed/12936957`

- **Fear of consequences.** Students may be dissatisfied with the food quality in the canteen, but refrain from reporting that to avoid price increase.

Summary of different types of bias is provided below:
Selection bias:

- Undercoverage
- Nonresponse
- Voluntary response

Response bias

- Non-neutral questions
- Non-anonymous surveys
- Fear of consequences

As it was said there could be present several types of bias in one survey. The Literary digest **survey** is an example of this. It is still important to distinguish between different types of biases to carefully avoid them. Otherwise you will not be able to prevent them.

Try not to be confused with *non-response*, *voluntary response* and *response* bias. While their names are almost the same, they are different things. Non-response and voluntary response biases are the subtypes of the selection bias. They are connected to the way an individual *gets into* the sample. An individual did not respond, so did not get into the sample, so this is a problem with selection. An individual wanted so much to respond, that he got into the sample when maybe he shouldn't. Contrary, the response bias is connected to the way an individual, who *is already in* the sample, gave his answer. It is a separate type of bias which includes several variants.

Once again, a large sample can *not* solve the methodological problems that produce bias. And again *Literary Digest example* discussed above illustrates this point. The sample size was over 2 millions respondents while it could not overcome the problems of the wrong sample.

Experiments

Experiment is a study where a researcher actively imposes some influence, called **treatment**, on a group of units, and then records the result. An experiment starts from identifying the point of interest. It is believed that some *explanatory variable* has an effect on a *response variable*. The researcher strives to determine and measure this effect. Participants of an experiment are called experimental units. Group that receives treatment is called treatment group.

Whenever possible it is necessary to add a **control group** to the experiment. It is a group that receives no treatment. It is needed to separate treatment from other uncontrolled possible forces that may arise during the experiment. The units in control group are left untouched during the study, but their results are measured. You can think of a control group as of a benchmark for better comparison.

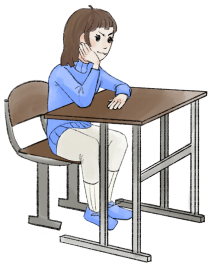
Without a control group, the level of response variable may change, but we will not be able to say whether it was changed on its own (by some other extraneous

variables) or by the treatment during the period of the study. For instance, a person might change a diet during the experiment, and the diet may result in a change of the response variable level. Such a change will not be reasoned by the treatment. Thus, the addition of a control group enables the researcher to assess the change due to the treatment alone, as opposed to the overall change.

So, all experimental units are divided into two groups, treatment and control.

Sometimes, treatment can be provided at different levels. They are then called treatment (or factor) levels. Or it could be several explanatory variables involved, like the amount of sun and the type of a fertilizer affecting the growth of plants. Then all the combinations used in the experiment will be called treatments

TO GO OR NOT TO GO. Masha wonders whether Statistics help rooms help to achieve higher scores on the final exams. She wants to conduct an experiment to check it. Then in such an experiment, the exam score will be the response variable and the help room-explanatory variable. Experimental units would be the students of the Statistics course. There would be students that would attend no classes to form a control group.



- explanatory variable (factor): *help room*
- response variable: *exam score*
- treatment group: *students attending help room*
- control group: *students not attending help room*

If we would like to introduce different levels of treatment that could be, for instance, 2h and 4h of help room per week.

Random assignment

Randomization in experiments means **random assignment** of subjects to treatments (or treatments to subjects which is the same). In Masha's experiment that would mean randomly assigning students to different hours of help room. In other words, we should divide experimental units into treatment and control groups randomly. The intuition behind is the following, the treatment and control groups should be "statistically equivalent" to each other with respect to subjects. This allows for equal initial conditions.

It must be always clear what is being randomized. Masha has to randomly assign which students out of her sample will attend what number of hours of help session per week, or she has to assign randomly for each student the amount of hours.

Recall that randomization, or random assignment, is a term that can only be used in the context of experiment. Contrary, when we talk about randomly choosing objects from population this term is inapplicable, but the term random sampling should be used. Make sure to distinguish between random sampling of units (from population into the sample to be studied) and randomized assignment to treatments. We usually do not speak about the former in experiments since in experiments our subjects are volunteers. In sampling we need randomness for representativeness of

the sample. In experiments we need randomness for compatibility of treatment and control groups.

Placebo effect and blinding

Sometimes, subjects may respond positively to any kind of treatment without a reason. Especially, this is common for medical experiments. Patients may feel better even though they are given a sugar pill, but told it is a strong medicine. This is called the placebo effect. To eliminate the *placebo effect* we use blinding - the procedure which ensures that subjects do not know whether they are in the treatment group or control group. The former get the treatment, while the latter get *placebo* - similarly looking empty/neutral treatment.

Double blinding occurs if those who collect data from patients also do not know who receives placebo and who receives real treatment. Double blinding may be necessary if the responses are evaluated subjectively, e.g. a physician assesses the general health condition of a patient based on his answers to questions. If the evaluation is objective, e. g. can be expressed in numbers, like blood pressure or heart rate, it may be not necessary to perform double blinding. Simple blinding is typically sufficient.

The logic is the following. We impose randomization to randomly assign subjects to treatment and control groups. But those who are in a treatment group would a priori be under the placebo effect. To eliminate it we artificially equalize the conditions and give placebos to control group. This does not make sense without a blinding.

Other variables interfering treatment effect

Confounding variables

Variables are said to be confounded if it is not clear which one causes the effect. Then confusion arises in an experiment which will lead to completely misleading results. It is caused by some **confounding variable** (confounding factor) which has also an effect on a response variable. As a result, an explanatory variable is confounded with such a factor. It is crucial to try to understand if confounding factors are present in the experiment.

Let's go back to the **"TO GO OR NOT TO GO"** example. Suppose there were 80 students involved in the experiment. Treatment was tested on group of 40 students and other 40 were in the control group. She used randomization properly. The following results were achieved:

Average exam score	
treatment	control
57	65

It is very peculiar – treatment group showed lower exam scores. How did this happen? It is a very desirable that fewer studies bring higher scores, but is it true?

Checking the experimental units more precisely, Masha found out the crucial detail. In control group it happened to be better students, who have *higher* exam scores by themselves. Contrary, in treatment group there were lazy students. So the difference in scores occured not due to the treatment! This happened because *students' activity* was an additional confounding factor, and then the *help session* and *students' activity* were confounded.

Lurking variables

Another factor that can create problems is a lurking variable. A **lurking variable** is an extraneous variable that affects both explanatory and response variables indicating a false relationship between them.

It could seem that the two variables are directly related to each other, while in fact they are not. This happens because some other third variable, lurking variable, affects them both. The presence of a lurking variable can lead to serious errors in the interpretation of results since they create misleading impression.

Recall the "SMART PEOPLE WEAR BIGGER SHOES" example from Chapter 3. The fact is that school pupils with larger shoe sizes appear to have higher reading levels. However, there exists another factor, *age*, which drives both variables. Elder students tend to have a bigger shoe size than younger students since they are more grown up, and elder pupils as well naturally have higher reading skills than elementary students. Age is a lurking variable here because it influences both the feet size (explanatory variable) and the intellect (response variable). Wearing larger shoes will not improve reading skills!

Such examples are "warning ups" not to infer causation from correlation.

Graphically the difference between lurking and confounding variables can be shown like in figure 1.3, where A is an explanatory variable, B is a response variable, and C is a lurking or a confounding variable correspondingly.

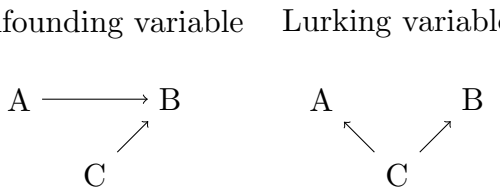


Figure 1.3: confounding and lurking variables

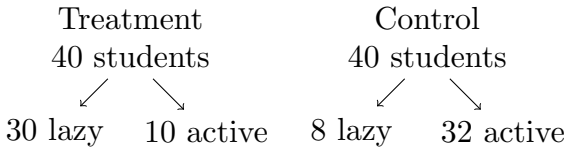
Blocking

To defeat this problem, the researcher can separate subjects into blocks, by a confounding factor. Then, each block will contain units with the same level of this confounding variable. This procedure is called blocking.

The common mistake that students make is they mix up blocking and *stratification*. They refer to different types of studies, one to an experimental and another to

an observational one, which are different in their essence. One may say that blocking is the analogue of stratification in experiments.

The actual distribution of active and lazy students in treatment and control groups was as following:



There were 38 lazy and 42 good students. Out of 38, 30 appeared to be in the treatment group, so the resulted average scores were low. Out of 42 active students, 32 were in the control group, so they showed really high results without any treatment.

If we look at all students divided by their activity in two groups, we will see different averages of exam scores. Now it is clear that the treatment provides better results.

Average exam score			
lazy students		active students	
treatment	control	treatment	control
51	48	75	69

Masha was right!
She didn't believe the misleading statement!
You have to study to achieve the higher marks.

Blocking versus randomization
Do not mix up blocking and randomization. First we do blocking, and after that we make randomization *within each block* by randomly assigning subjects to treatment and control groups. In experiments "group" is not a pure synonym of "block". Block is mostly a formal term. We use blocking by a certain confounding factor obvious to us, and we use randomization to reduce effects of other potential confounding factors unavailable to our mind. The logic is the following. We fix one confounding factor which is clear to us and make blocking by it. To fix other possible confounding factors which are not clear to us, we make randomization so that in treatment in control groups would be different units.

Types of Experimental Designs

According to procedures described above there are different types of experimental design: **completely randomized** design, when there is no blocking, and **randomized block** design, when blocking is used.

There is also a special case of blocking – **matched pairs** design (or paired comparison design), in which each pair can be considered as a block. The crucial fact in this type is that treatment is done on the *same* units at different periods of time

or circumstances. For example, a study is conducted in a big company and we are checking the level of work satisfaction of the same managers before and after the crisis.

Overall an experiment can be shown with a diagram like the one below. Here we have an example of a randomized block design. Before understanding that active and lazy students exist, Masha had a completely randomized design.

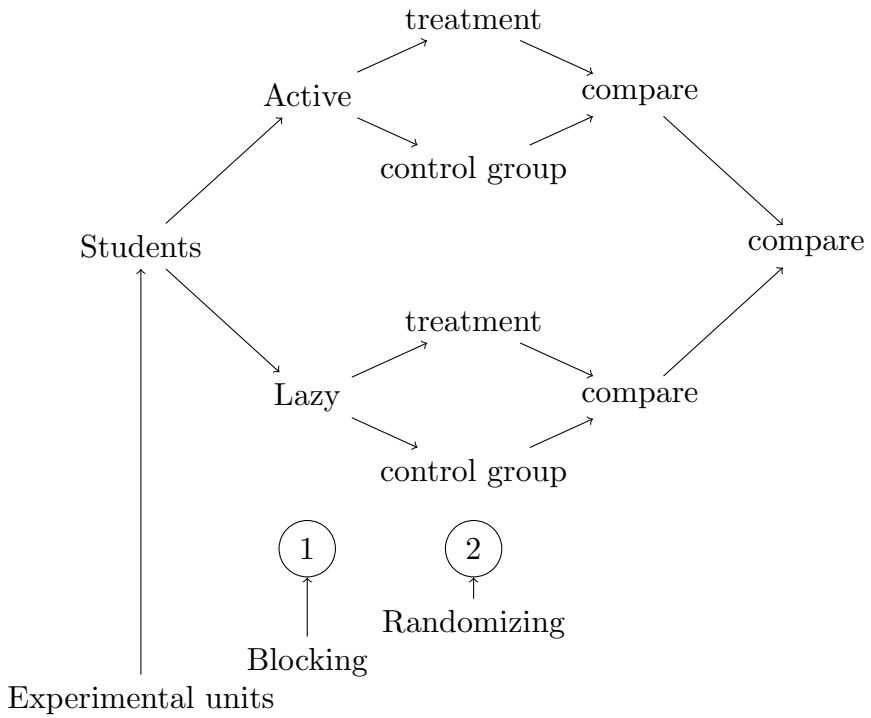


Figure 1.4: Scheme of Masha's experiment "To go or not to go?"

Note that there are two levels of comparison. Based on the structure of experiment, blinding is unfortunately not possible here.

Principles of good experiment

Every good experiment must satisfy the following criteria:

1. **Control** – who receives what treatments, conditions should be as similar as possible for all involved groups

2. **Replication** – the possibility to repeat an experiment many times to confirm that the obtained result is stable.

   Some differences could be obtained just by chance (it happened this time but may not happen next time). Then, the obtained relation is insignificant. One important factor is also the size of the sample since the difference could be explained by the natural variation – the larger the sample is, the more significant is the observation. The treatment should be repeated on a sufficient number of subjects so that the real differences are noted.

Why do we want an experiment to be repeated?

(a) We want to conduct an experiment in different environment/country.

(b) We want to conduct an experiment over a period of time.

(c) If we do not trust the researcher and want to check by our own.

To understand the principle of replication you can think of a good experiment as of a good *recipe* of meal: another person has to be able to cook the same meal, in another country, 10 years later.

3. **Generalizability** – the possibility to apply the results of an experiment to other sets of experimental units, that is, to *generalize* the results for a wider range. In our experiment, if you discover the effect of help rooms on this year students, you want the results you obtained to be applicable, for instance, to the next year students or to the students of other courses. Sometimes it is not possible. For instance, it is hard to generalize from the effect of a TV advertisement on high school students to the effect of the same advertisement on retired senior citizens.

As you've seen there are many tricky factors in every experiment. In order to get reliable results the experiment should be conducted properly. A researcher should control many factors!

Comparison of Observational Study and Experiment

By definition, experiment implies controlled assignment of treatments to subjects. Observational study does not imply this.

Experiment	Assign influence (treatment)
Observational Study	Try to make no presence

In experiment the researcher imposes and controls the treatment, and thus can conclude on causal relationship (what causes what). In other words, in experiment we can test whether treatment has an effect on the response variable. In observational study the researcher cannot determine which variables affect the response. However while results may suggest relationships, it is impossible to conclude the cause and effect relationship. This is why observational studies may show only the existence of *associations/relations/tendencies* but not assure causality.

As a result of the differences in design of experiment and observation studies, different conclusions can be inferred from them.

Type of study	Conclusions that may be inferred
Experiment	Cause-and-effect relationships
Observational study	Only associations. No statements about causality

Full score strategy

Experiment: Random assignment to treatment.

1. Give integers to all experimental units in ascending order.
2. Using a random number generator, generate the required for treatment group number of integers, ignoring repeats and out of range.
3. Place the experimental units corresponding to generated integers in the treatment group.
4. The rest will go to control group.

This is a weak place of many students. They do not think that so exact steps are required. However they are needed, otherwise the process of random assignment will not be described fully. You may think that you are writing a program of randomization for a computer.

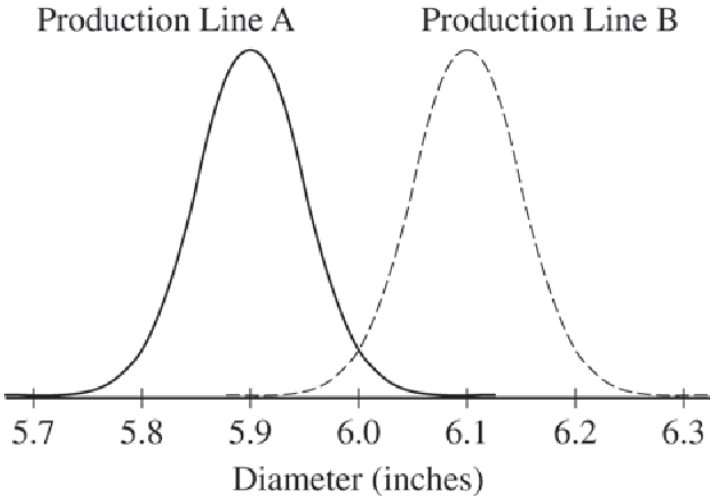
You must be able to reproduce even being drunk:

- how to distinguish between observational study and experiment
- describe the procedure of SRS and why it is crucial
- sampling error is present in every sample survey, it is not a threat for credibility of results
- bias is a threat for credibility because it systematically gives errors
- explanatory variable, response variable, treatment and control groups
- random assignment of experimental units
- placebo effect and blinding
- blocking to defeat confounding factors

Sample AP problems with solutions

Problem 1. AP 2015 №6 a

Corn tortillas are made at a large facility (завод) that produces 100,000 tortillas per day on each of its two production lines. The distribution of the diameters of the tortillas produced on production line A is approximately normal with mean 5.9 inches, and the distribution of the diameters of the tortillas produced on production line B is approximately normal with mean 6.1 inches. The figure below shows the distributions of diameters for the two production lines.



The tortillas produced at the factory are advertised as having a diameter of 6 inches. For the purpose of quality control, a sample of 200 tortillas is selected and the diameters are measured. From the sample of 200 tortillas, the manager of the facility wants to estimate the mean diameter, in inches, of the 200,000 tortillas produced on a given day. Two sampling methods have been proposed.

- *Method 1:* Take a random sample of 200 tortillas from the 200,000 tortillas produced on a given day. Measure the diameter of each selected tortilla.
- *Method 2:* Randomly select one of the two production lines on a given day. Take a random sample of 200 tortillas from the 100,000 tortillas produced by the selected production line. Measure the diameter of each selected tortilla.

(a) Will a sample obtained using Method 2 be representative of the population of all tortillas made that day, with respect to the diameters of the tortillas? Explain why or why not.

Solution

(a) No, a sample obtained using Method 2 would not be representative of all tortillas made that day, with respect to diameter.

This happens because the sample would represent the tortillas only from one production line, and not from the entire population. The sample would not represent the entire populations since the tortillas from the different produc-

tion lines are different with respect to diameter.

Problem 2. AP 2014 №2 c Nine sales representatives, 6 men and 3 women, at a small company wanted to attend a national convention. There were only enough travel funds to send 3 people. The manager selected 3 people to attend and stated that the people were selected at random. The 3 people selected were women. There were concerns that no men were selected to attend the convention.

(c) An alternative to calculating the exact probability that randomly selecting 3 people from a group of 6 men and 3 women will result in selecting 3 women is to conduct a simulation to estimate the probability. A proposed simulation process is described below.

> Each trial in the simulation consists of rolling three fair, six sided dice, one die for each of the convention attendees. For each die, rolling a 1, 2, 3, or 4 represents selecting a man; rolling a 5 or 6 represents selecting a woman. After 1,000 trials, the number of times the dice indicate selecting 3 women is recorded.

Does the proposed process correctly simulate the random selection of 3 women from a group of 9 people consisting of 6 men and 3 women? Explain why or why not.

Solution
No, the process does not correctly simulate the random selection of three women from a group of nine people, of whom six are men and three are women.

Explanation:

The dice outcomes in the proposed simulation are independent while the genders of the selected attendees are dependent.

OR

In the simulation with the dice, the three dice rolls in any given trial are independent of one another, indicating a selection process that is done *with* replacement. However, the random selection of three people among nine is done *without* replacement

OR

The probability of rolling a 5 or a 6 is the same on all three dice while the probability of selecting a woman changes after each selection.

Problem 3. AP 2014 №4 b As part of its twenty-fifth reunion celebration, the class of 1988 (students who graduated in 1988) at a state university held a reception on campus. In an informal survey, the director of alumni development asked 50 of the attendees about their incomes. The director computed the mean income of the 50 attendees to be $189,952. In a news release, the director announced, "The members of our class of 1988 enjoyed resounding success. Last year's mean income of its members was $189,952!"

(b) The director felt the members who attended the reception may be different from the class as a whole. A more detailed survey of the class was planned to find a better estimate of the income as well as other facts about the alumni. The staff developed two methods based on the available funds to carry out the survey.

- *Method 1:* Send out an e-mail to all 6,826 members of the class asking them to complete an online form. The staff estimates that at least 600 members will respond.
- *Method 2:* Select a simple random sample of members of the class and contact the selected members directly by phone. Follow up to ensure that all responses are obtained. Because method 2 will require more time than method 1, the staff estimates that only 100 members of the class could be contacted using method 2.

Which of the two methods would you select for estimating the average yearly income of all 6,826 members of the class of 1988 ? Explain your reasoning by comparing the two methods and the effect of each method on the estimate.

Solution:

1. Choose method *and* identify relevant characteristic for each method.
   Indicate the effect of the biased method.
   Income of responders may differ from income of non-responders.
   Biased method produce misleading estimate of the mean income, including direction.
   Method 2 is better than Method 1.
   A sample obtained from Method 1 could be *biased* because of the non-response bias since the expected response rate is very low (only few people would respond).
   With Method 2, despite the smaller sample size, the *random selection* is likely to result in a sample that is more *representative* of the entire population (class) and produce an *unbiased estimate* of mean yearly income of all class members.
   It is plausible that class members *with larger incomes* might be *more likely to return* the form than class members with smaller incomes. (rich people are more likely to respond).
   The mean income for such a sample would *overestimate* the mean income of all class members.

Problem 4. AP 2013 №2 An administrator at a large university wants to conduct a survey to estimate the proportion of students who are satisfied with the appearance of the university buildings and grounds. The administrator is considering three methods of obtaining a sample of 500 students from the 70,000 students at the university.

1. Because of financial constraints, the first method the administrator is considering consists of taking a convenience sample to keep the expenses low. A very

large number of students will attend the first football game of the season, and the first 500 students who enter the football stadium could be used as a sample. Why might such a sampling method be biased in producing an estimate of the proportion of students who are satisfied with the appearance of the buildings and grounds?

2. Because of the large number of students at the university, the second method the administrator is considering consists of using a computer with a random number generator to select a simple random sample of 500 students from a list of 70,000 student names. Describe how to implement such a method.

3. Because stratification can often provide a more precise estimate than a simple random sample, the third method the administrator is considering consists of selecting a stratified random sample of 500 students. The university has two campuses with male and female students at each campus. Under what circumstance(s) would stratification by campus provide a more precise estimate of the proportion of students who are satisfied with the appearance of the university buildings and grounds than stratification by gender?

Solution:

(a) The first 500 students who enter the football stadium were not likely to be representative of the population of all students at the university. In other words, these 500 students were likely to differ systematically from the population with regard to many variables. For example, these 500 students might have more school pride than the population of students as a whole, which might be related to their opinions about the appearance of university buildings and grounds. Perhaps their school pride is related to having more positive opinions about the appearance of university buildings and grounds, in which case the sample proportion of students who were satisfied would be biased toward overestimating the population proportion of students who were satisfied.

(b) Obtain a list of all 70,000 students at the university. Assign an identification number from 1 to 70,000 to each student. Then use a computer to generate 500 random integers between 1 and 70,000 without replacement. The students whose ID numbers correspond to those numbers were then selected for the sample.

(c) Stratifying by campus would be more advantageous than stratifying by gender provided that opinions about appearance of university buildings and grounds between the two campuses differ more than the opinions about appearance of university buildings and grounds between the two genders.

Problem 5. AP 2012 №5 c

A recent report stated that less than 35 percent of the adult residents in a certain city will be able to pass a physical fitness test. Consequently, the city's Recreation

Department is trying to convince the City Council to fund more physical fitness programs. The council is facing budget constraints and is skeptical of the report. The council will fund more physical fitness programs only if the Recreation Department can provide convincing evidence that the report is true.

The Recreation Department plans to collect data from a sample of 185 adult residents in the city. It recruits 185 adult residents who volunteer to take the physical fitness test. The test is passed by 77 of the 185 volunteers. Describe the primary flaw in this study, and explain why it is a concern.

Solution:
This is not a randomly selected sample because the sample was selected by recruiting volunteers. It seems reasonable to think that volunteers would be more physically fit than the population of city adults as a whole. Therefore, the sample proportion will likely *overestimate* the population proportion of adult residents in the city who are able to pass the physical fitness test.

Problem 6. AP 2012 №6 a Two students at a large high school, Peter and Rania, wanted to estimate m, the mean number of soft drinks that a student at their school consumes in a week. A complete roster (перечень) of the names and genders for the 2,000 students at their school was available. Peter selected a simple random sample of 100 students. Rania, knowing that 60 percent of the students at the school are female, selected a simple random sample of 60 females and an independent simple random sample of 40 males. Both asked all of the students in their samples how many soft drinks they typically consume in a week.

(a) Describe a method Peter could have used to select a simple random sample of 100 students from the school.

Solution:
Peter can *number* the students from 1 to 2,000 and then use a calculator or computer to *generate* 100 unique random numbers between 1 and 2,000 *without replacement*. If non-unique numbers are generated, the *repeated* numbers are ignored until 100 unique numbers are obtained. The students whose numbers *correspond* to the randomly generated numbers are then selected for the sample.

Problem 7. AP 2011B №2
People with acrophobia (fear of heights) sometimes enroll in therapy sessions to help them overcome this fear. Typically, seven or eight therapy sessions are needed before improvement is noticed. A study was conducted to determine whether the drug D-cycloserine, used in combination with fewer therapy sessions, would help people with acrophobia overcome this fear.

Each of 27 people who participated in the study received a pill before each of two therapy sessions. Seventeen of the 27 people were randomly assigned to receive a D-cycloserine pill, and the remaining 10 people received a placebo. After the two therapy sessions, none of the 27 people received additional pills or therapy. Three months after the administration of the pills and the two therapy sessions, each of the 27 people was evaluated to see if he or she had improved.

(a) Was this study an experiment or an observational study? Provide an explanation to support your answer.

(b) When the data were analyzed, the D-cycloserine group showed statistically significantly more improvement than the placebo group did. Based on this result, would the researchers be justified in concluding that the D-cycloserine pill and two therapy sessions are as beneficial as eight therapy sessions without the pill? Justify your answer.

(c) A newspaper article that summarized the results of this study did not explain how it was determined which people received D-cycloserine and which received the placebo. Suppose the researchers allowed the therapists to choose which people received D-cycloserine and which received the placebo, and no randomization was used. Explain why such a method of assignment might lead to an incorrect conclusion.

Solution:

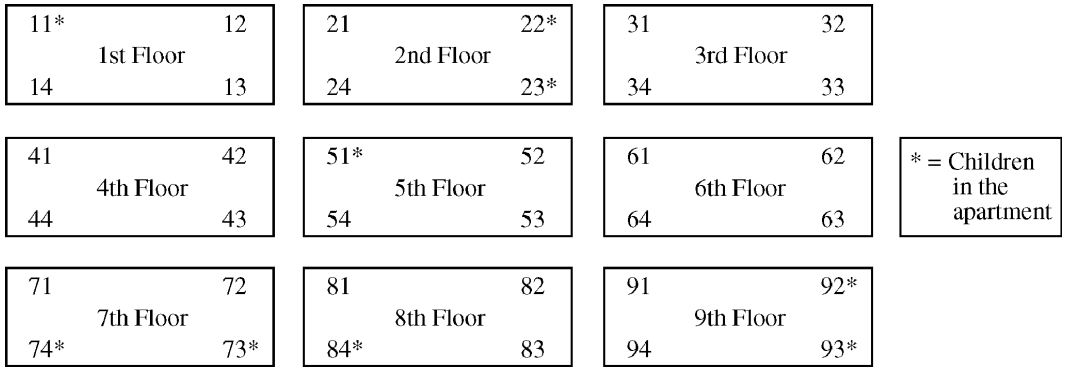
(a) The study was an experiment because treatments (D-cycloserine or placebo) were imposed by the researchers on the people with acrophobia.

(b) No, the experiment was designed to compare the D-cycloserine group with a control group that received the placebo. The researchers can conclude that the D-cycloserine pill and two therapy sessions show significantly more improvement than a placebo and two therapy sessions. However, there is no basis for comparison with another group of people with acrophobia who received eight therapy sessions and no pill.

(c) One example is that if the therapists were allowed to choose who received the placebo and who received D-cycloserine, they might assign the people with more severe acrophobia to one of the groups and the people with less severe acrophobia to the other group. Thus, the improvement after only two therapy sessions could be related to the initial severity of the acrophobia rather than to the effects of D-cycloserine.

Problem 8. AP 2011 №3 An apartment building has nine floors and each floor has four apartments. The building owner wants to install new carpeting in eight apartments to see how well it wears before she decides whether to replace the carpet in the entire building.

The figure below shows the floors of apartments in the building with their apartment numbers. Only the nine apartments indicated with an asterisk (*) have children in the apartment.



(a) For convenience, the apartment building owner wants to use a cluster sampling method, in which the floors are clusters, to select the eight apartments. Describe a process for randomly selecting eight different apartments using this method.

(b) An alternative sampling method would be to select a stratified random sample of eight apartments, where the strata are apartments with children and apartments with no children. A stratified random sample of size eight might include two randomly selected apartments with children and six randomly selected apartments with no children. In the context of this situation, give one statistical advantage of selecting such a stratified sample as opposed to a cluster sample of eight apartments using the floors as clusters.

Solution:

(a) The following two-step process can be used to select the eight apartments.

Step 1: Generate a random integer between 1 and 9, inclusive, using a calculator, a computer program, or a table of random digits. Select all four apartments on the floor corresponding to the selected integer.

Step 2: Generate another random integer between 1 and 9, inclusive. If the generated integer is the same as the integer generated in step 1, continue generating random integers between 1 and 9 until a different integer appears. Again select all four apartments on the floor corresponding to the second selected integer.

The cluster sample consists of the eight apartments on the two randomly selected floors.

(b) Because the amount of wear on the carpets in apartments with children could be different from the wear on the carpets in apartments without children, it would be advantageous to have apartments with children represented in the sample. The cluster sampling procedure in part (a) could produce a sample with no children in the selected apartments; for example, a cluster sample of the apartments on the third and sixth floors would consist entirely of apartments with no children. Stratified random sampling, where the two strata are apartments with children and apartments without children, guarantees a

sample that includes apartments with and without children, which, in turn, would yield sample data that are representative of both types of apartments.

Problem 9. AP 2010 Form B №2

In response to nutrition concerns raised last year about food served in school cafeterias, the Smallville School District entered into a one-year contract with the Healthy Alternative Meals (HAM) company. Under this contract, the company plans and prepares meals for 2,500 elementary, middle, and high school students, with a focus on good nutrition.

The school administration would like to survey the students in the district to estimate the proportion of students who are satisfied with the food under this contract. Two sampling plans for selecting the students to be surveyed are under consideration by the administration. One plan is to take a simple random sample of students in the district and then survey those students. The other plan is to take a stratified random sample of students in the district and then survey those students.

(a) Describe a simple random sampling procedure that the administrators could use to select 200 students from the 2,500 students in the district.

(b) If a stratified random sampling procedure is used, give one example of an effective variable on which to stratify in this survey. Explain your reasoning.

(c) Describe one statistical advantage of using a stratified random sample over a simple random sample in the context of this study.

Solution:s

(a) The administrators could number an alphabetical list of students from 1 to 2,500. They could then use a random number generator from a calculator or computer to generate 200 unique random integers from 1 to 2,500.The students corresponding to those 200 numbers would be asked to participate in the survey. Question assesses students' ability to describe a simple random sampling procedure.

(b) One possible stratification variable might be the school level of the student (elementary, middle, high school). The students' perceptions of the importance of good nutrition in food served may differ depending on the students' ages and therefore on school levels. For example, there may be a difference between what elementary students value in food served as opposed to middle school and high school students.

(c) One statistical advantage of using stratified random sampling as opposed to simple random sampling is the following. For example, elementary, middle and high school strata create groups that differ with respect to opinion on food satisfaction. Therefore, the same overall sample size will produce more accurate estimate of the population proportion of students who are satisfied with the food.

Another advantage is that stratified random sampling guarantees that each of the school-level strata will have some representation, because it is possible that a simple random sample would miss one or more of the strata completely.

Problem 10. AP 2010 №1 a

Agricultural experts are trying to develop a bird deterrent to reduce costly damage to crops in the United States. An experiment is to be conducted using garlic oil to study its effectiveness as a nontoxic, environmentally safe bird repellant. The experiment will use European starlings, a bird species that causes considerable damage annually to the corn crop in the United States. Food granules made from corn are to be infused with garlic oil in each of five concentrations of garlic – 0 percent, 2 percent, 10 percent, 25 percent, and 50 percent. The researchers will determine the adverse reaction of the birds to the repellant by measuring the number of food granules consumed during a two-hour period following overnight food deprivation. There are forty birds available for the experiment, and the researchers will use eight birds for each concentration of garlic. Each bird will be kept in a separate cage and provided with the same number of food granules.

(a) For the experiment, identify

i. the treatments

ii. the experimental units

iii. the response that will be measured

Solution:

i. The treatments are the different concentrations of garlic in the food granules. Specifically, there are five treatments: 0 percent, 2 percent, 10 percent, 25 percent and 50 percent.

ii. The experimental units are the birds (starlings), each placed in an individual cage.

iii. The response is the number of food granules consumed by the bird.

Problem 11. AP 2009 №3

Before beginning a unit on frog anatomy, a seventh-grade biology teacher gives each of the 24 students in the class a pretest to assess their knowledge of frog anatomy. The teacher wants to compare the effectiveness of an instructional program in which students physically dissect frogs with the effectiveness of a different program in which students use computer software that only simulates the dissection of a frog. After completing one of the two programs, students will be given a posttest to assess their knowledge of frog anatomy. The teacher will then analyze the changes in the test scores (score on posttest minus score on pretest).

(a) Describe a method for assigning the 24 students to two groups of equal size that allows for a statistically valid comparison of the two instructional programs.

(b) Suppose the teacher decided to allow the students in the class to select which instructional program on frog anatomy (physical dissection or computer simulation) they prefer to take, and 11 students choose actual dissection and 13 students choose computer simulation. How might that self-selection process jeopardize a statistically valid comparison of the changes in the test scores (score on posttest minus score on pretest) for the two instructional programs? Provide a specific example to support your answer.

Solution:

(a) There are several possible answers.

Method 1. **Completely randomized design.**

Each student will be assigned a unique random number using a random number generator on a calculator, statistical software, or a random number table. The assigned numbers will be listed in ascending order. The students with the lowest 12 numbers in the ordered list will receive the instructional program that requires physically dissecting frogs. The students with the highest 12 numbers will receive the instructional program that uses computer software to simulate the dissection of a frog.

Method 2. **Randomized Block design.**

Students will be paired or placed into blocks of size two, based on having similar pretest scores. So, the first block will contain the two students with the two lowest pretest scores, the second block will contain the two students with the third- and fourth-lowest pretest scores, and so on, with the last block containing the two students with the two highest pretest scores. In each block, the students will be assigned a unique random number using a random number generator on a calculator, statistical software, or a random number table. The student in each block with the lower random number will receive the instructional program that requires physically dissecting frogs, and the student with the higher random number will receive the instructional program that uses computer software to simulate the dissection of a frog.

(b) By not randomizing and allowing the students to self-select, there is a potential for changes to occur in the differences between pretest and posttest scores for a particular group because of the characteristics of students who choose a particular instructional method, not because of the instructional method itself. For example, suppose frog-loving students already know a lot about frog anatomy; one would therefore expect these students to be less likely to show a large change between the pretest and posttest scores. Suppose the frog-loving students tend to select the computer simulation method (perhaps because they do not like the notion of dissecting the frogs they love). The possible low change between pretest and posttest scores for the computer

simulation group might then be attributed to the students' already knowing a lot about frog anatomy beforehand, not to the instructional method itself. The frog dissection group might see a larger change in scores because the students entering this group are those with the lower pretest scores (less prior knowledge) and who are thus more likely to show greater improvement between pretest and posttest scores.

Problem 12. AP 2009B №6 a

Two treatments, A and B, showed promise for treating a potentially fatal disease. A randomized experiment was conducted to determine whether there is a significant difference in the survival rate between patients who receive treatment A and those who receive treatment B. Of 154 patients who received treatment A, 38 survived for at least 15 years, whereas 16 of the 164 patients who received treatment B survived at least 15 years.

(a) Treatment A can be administered only as a pill, and treatment B can be administered only as an injection. Can this randomized experiment be performed as a double-blind experiment? Why or why not?

Solution:

Yes, it can. Note that big peace of information like "one treatment could be administered only as a pill and another.., etc" you can consider as additional and is given to confuse student. It is connected with the medicine and not at all with statistical course.

Answer is that an experiment is said to be double-blind when the control or comparison group and the treatment group are treated exactly in the same manner *(except, of course, for the levels of treatments of interest*) so that neither the patients nor the researchers who care for the patients and take measurements know which treatment has been assigned to which patient.

Because treatment A and treatment B are administered differently, then a placebo for each treatment group could be introduced to make the study double-blind. Namely, the patients who get treatment A also get a placebo injection and the patients who get treatment B also get a placebo pill. The physician who treats patients does not have to know which treatment the patient is receiving. Treatment plus placebo is sent to physician by a researcher not in contact with the patients (this researcher is the only one who knows which treatment each patient is receiving).

Problem 13. AP 2008 №2

A local school board plans to conduct a survey of parents' opinions about yearround schooling in elementary schools. The school board obtains a list of all families in the district with at least one child in an elementary school and sends the survey to a random sample of 500 of the families. The survey question is provided below.

A proposal has been submitted that would require students in elementary schools to attend school on a yearround basis. Do you support this proposal? (Yes or No)

The school board received responses from 98 of the families, with 76 of the responses indicating support for year-round schools. Based on this outcome, the local school board concludes that most of the families with at least one child in elementary school prefer year-round schooling.

(a) What is a possible consequence of nonresponse bias for interpreting the results of this survey?

(b) Someone advised the local school board to take an additional random sample of 500 families and to use the combined results to make their decision. Would this be a suitable solution to the issue raised in part (a)?

Explain.

(c) Suggest a different follow-up step from the one suggested in part (b) that the local school board could take to address the issue raised in part (a).

Solution:

(a) Responses were received from only 98 of the 500 (or 19.6 percent) of the randomly selected families. In other words, 80.4 percent of the randomly selected families did not respond to the survey. To obtain a nearly unbiased estimate of the proportion of families with at least one child in elementary school in this school district who support year-round school, we would need to assume that the families that did not respond would have a similar level of support for year-round school as those who did respond. This would not be the case, for example, if families who support year-round school were more likely to respond than families who do not support year-round school. In such a case, the estimate of the proportion of families who support year-round school calculated from the responses would tend to be higher than the population proportion of families who favor year-round school.

(b) No, the nonresponse bias still exists. Combining the results from the original sample with a new random sample of 500 families will not solve the problem. Regardless of what happens in the second sample, the problem of nonresponse bias will still exist in the combined sample because there would be at least 402 nonresponses included from the original sample.

Contact the 402 families from whom responses were not received and ask their opinion on the proposal. This may require additional mailings or telephone calls, but it will provide better information about support for year-round school among all families in the school district with at least one child in elementary school.

OR

Take a new random sample or take a census and use an alternative strategy, such as telephone calls or inperson interviews, to help increase the response rate.

Problem 14. AP 2008 Form B №4 a

A researcher wants to conduct a study to test whether listening to soothing music for 20 minutes helps to reduce diastolic blood pressure in patients with high blood pressure, compared to simply sitting quietly in a noise-free environment for 20 minutes. One hundred patients with high blood pressure at a large medical clinic are available to participate in this study.

(a) Propose a design for this study to compare these two treatments.

Solution:

(a) There are several ways to answer this part. You can choose either of them which you like more or understand better.

Approach 1: Paired Design

Each subject will receive both treatments, with a suitable length of time between treatments. The order of the treatments will be randomly assigned to the subjects. For example, for each patient flip a coin to determine which treatment will be administered first. Measure diastolic blood pressure, then have the subject sit quietly for 20 minutes in either a noise-free environment or in a room where soothing music is played, depending on which treatment was selected at random (based on the coin flip). At the end of the 20 minutes, measure diastolic blood pressure again and compute its change (*after − before).* After a suitable period of time, repeat with the other treatment.

When the data have been collected, the difference (*music − noise-free*) in the change in diastolic blood

pressure will be computed for each subject, and then a paired t-test will be run to see if the mean difference is significantly different from zero.

Approach 2: Matched Pairs Design

Measure diastolic blood pressure for each of the 100 subjects and then form 50 pairs based on these readings by pairing the two with the highest diastolic blood pressure, then the two with the next highest, and so on. For each pair, toss a coin to determine which member of the pair will be assigned to group 1, and then assign the other member of the pair to group 2. For group 1, measure diastolic blood pressure, then have the subjects sit quietly in a noise-free environment for 20 minutes, and then measure diastolic blood pressure again and compute its change (*after − before).* For group 2, the plan is the same, except that they will sit for 20 minutes in a room where soothing music is played between blood pressure measurements.

When the data have been collected, the difference (*music − noise-free*) in the change in diastolic blood

pressure will be computed for each pair, and then a paired t-test will be run to see if the mean difference is significantly different from zero.

Approach 3: Completely Randomized Design (This is not as good a choice as the two previous approaches, but because of the large number of subjects available for each treatment group, it is considered an acceptable solution.)

Assign the 100 patients numbers from 00 to 99. From a random number table, select 50 unique numbers; the patients with the selected values will form group 1; the remaining 50 patients will form group 2. For group 1, measure diastolic blood pressure, then have the subjects sit quietly in a noise-free environment for 20 minutes, and then measure diastolic blood pressure again. For group 2, the plan is the same, except that they will sit for 20 minutes in a room where soothing music is played between blood pressure measurements. When the data have been collected, the change in diastolic blood pressure will be computed for each subject, and then a two-sample t-test will be run to see if there is a significant difference between the mean change attributable to *music* and the mean change attributable to a *noise-free* environment.

Problem 15. AP 2007 №5 a

Researchers want to determine whether drivers are significantly more distracted while driving when using a cell phone than when talking to a passenger in the car. In a study involving 48 people, 24 people were randomly assigned to drive in a driving simulator while using a cell phone, while the remaining 24 were assigned to drive in the driving simulator while talking to a passenger in the simulator. Part of the driving simulation for both groups involved asking drivers to exit the freeway at a particular exit. In the study, 7 of the 24 cell phone users missed the exit and 2 of the 24 talking to a passenger missed the exit.

Would you classify this study as an experiment or an observational study? Provide an explanation to support your answer.

Solution:
This is clearly an experiment. The researchers imposed treatments by randomly assigning drivers to the two different categories (1-simulated driving while talking on a cell phone and 2-simulated driving while talking to a passenger).

Problem 16. AP 2006 №5

A biologist is interested in studying the effect of growth-enhancing nutrients and different salinity (salt) levels in water on the growth of shrimps. The biologist has ordered a large shipment of young tiger shrimps from a supply house for use in the study. The experiment is to be conducted in a laboratory where 10 tiger shrimps are placed randomly into each of 12 similar tanks in a controlled environment. The biologist is planning to use 3 different growth-enhancing nutrients (A, B, and C) and two different salinity levels (low and high).

(a) List the treatments that the biologist plans to use in this experiment.

(b) Using the treatments listed in part (a), describe a completely randomized design that will allow the biologist to compare the shrimps' growth after 3 weeks.

(c) Give one statistical advantage to having only tiger shrimps in the experiment. Explain why this is an advantage.

(d) Give one statistical disadvantage to having only tiger shrimps in the experiment. Explain why this is a disadvantage.

Solution:

1. The three different growth-enhancing nutrients (A, B, and C) and two different salinity levels (low and high) yield a total of 6 different treatment combinations for this experiment: A-low, A-high, B-low, B-high, C-low, C-high.

2. Since 10 tiger shrimps have already been randomly placed into each of 12 similar tanks in a controlled environment, we must randomly assign the treatment combinations to the tanks. Each treatment combination will be randomly assigned to 2 of the 12 tanks. One way to do this is to number the tanks from 1 to 12, then arrange numbers from 1 to 12 in a random order, then assign treatment 1 to tanks with numbers equal to the first and second random number; assign treatment 2 to the tanks with numbers equal to the third and fourth random numbers, and so on.

3. Using only tiger shrimp will reduce a source of variation in the experimental units, the tanks of shrimp in this experiment. By eliminating this possible source of variation, type of shrimp, we are better able to isolate the variability due to the factors of interest to us (nutrient and salinity level). This will make it easier to identify any treatment effects that may be present.

4. Using only tiger shrimp will limit the scope of inference for the biologist. Ideally, the biologist would like to identify the treatment combination that leads to the most growth for all shrimp. However, the biologist will only be able to identify the best treatment combination for tiger shrimp because other types of shrimp may respond differently to the treatments.

Problem 17. AP 2006 Form B №5

When a tractor pulls a plow (плуг) through an agricultural field, the energy needed to pull that plow is called the draft. The draft is affected by environmental conditions such as soil type, terrain, and moisture.

A study was conducted to determine whether a newly developed hitch (крепеж) would be able to reduce the draft compared to the standard hitch. (A hitch is used to connect the plow to the tractor) Two large plots of land were used in this study. It was randomly determined which plot was to be plowed using the standard hitch. As the tractor plowed that plot, a measurement device on the tractor automatically recorded the draft at 25 different randomly selected points of the plot.

After the plot was plowed, the hitch was changed from the standard one to the new one, a process that takes a substantial amount of time. Then the second plot was plowed using the new hitch. Twenty-five measurements of draft were also recorded at randomly selected points in this plot.

1. What was the response variable in this study?

   Identify the treatments.

   What were the experimental units?

2. Given that the goal of the study is to determine whether a newly developed hitch reduces draft compared to the standard hitch, was randomization used properly in this study? Justify your answer.

3. Given that the goal of the study is to determine whether a newly developed hitch reduces draft compared to the standard hitch, was replication used properly in this study? Justify your answer.

4. Plot of land is a confounding variable in this experiment. Explain why.

Solution:

(a) The response variable was the draft, namely the amount of draft. The two treatments were the standard hitch and the new hitch. The experimental units were the two large plots of land.

(b) Yes, it was used properly. The two hitches (treatments) were randomly assigned to the two plots (experimental units).

(c) No, it wasn't because each treatment (type of hitch) was applied to only one experimental unit (plot of land). Replication is used to repeat the treatments on different experimental units so general pattern could be observed. In this study there was no replication.

(d) (Here they mean the possible different conditions of two plots were meant)

Even though 25 measurements were taken at different locations in the two plots, each hitch was used in one plot only. Thus, if any difference in the draft is finally observed we will not be sure whether the difference is due to the hitch or the plot (moisture, quality of ground, etc.) That means plot could be a confounding variable and the treatments (hitches) are confounded with the plots.

P.S. note that here it happened that the experimental units are at the same time confounding factors.

Problem 18. AP 2005 №1 b,c

The goal of a nutritional study was to compare the caloric intake of adolescents (молодой человек, юноша) living in rural areas of the United States with the caloric intake of adolescents living in urban areas of the United States. A random sample of ninth-grade students from one high school in a rural area was selected. Another random sample of ninth graders from one high school in an urban area was also selected. Each student in each sample kept records of all the food he or she consumed in one day.

(b) Is it reasonable to generalize the findings of this study to all rural and urban ninth-grade students in the United States? Explain.

(c) Researchers who want to conduct a similar study are debating which of the following two plants to use.

Plan I. Have each student in the study record all the food he or she consumed in one day. Then researchers would compute the number of calories of food consume per kilogram of body weight for each student for that day.

Plan II. Have each student in the study record all the food he or she consumed over the same 7-day period. Then researchers would compute the average daily number of calories of food consume per kilogram of body weight for each student during that 7-day period.

Assuming that the students keep accurate records, which plan, I or II, would better meet the goal of the study? Justify your answer.

Solution:

1. No. The samples include students only from one rural and one urban school, so it is not reasonable to generalize findings to all of the school in the United States. For that you need to conduct an appropriate sampling procedure from all of the schools of the United States.

2. Plan II will do a better job of comparing a daily caloric intake. Both plans take body weight onto account, but plan II takes 7 days into account. There are possibly differences in caloric intake among days, so it would be better to average over the 7 days than to take only one day into consideration. Therefore, plan II provides with a better

Problem 19. AP 2005 Form B №3

In search of a mosquito repellent that is safer than the ones that are currently on the market, scientists have developed a new compound that is rated as less toxic than the current compound, thus making a repellent that contains this new compound safer for human use. Scientists also believe that a repellent containing the new compound will be more effective than the ones that contain the current compound. To test the effectiveness of the new compound versus that of the current compound, scientists have randomly selected 100 people from a state. Up to 100 bins, with an equal number of mosquitoes in each bin, are available for use in the study. After a compound is

applied to a participant's forearm, the participant will insert his or her forearm into a bin for 1 minute, and the number of mosquito bites on the arm at the end of that time will be determined.

(a) Suppose this study is to be conducted using a completely randomized design. Describe a randomization process and identify an inference procedure for the study.

(b) Suppose this study is to be conducted using a matched-pairs design. Describe a randomization process and identify an inference procedure for the study.

(c) Which of the designs, the one in part (a) or the one in part (b), is better for testing the effectiveness of the new compound versus that of the current compound? Justify your answer.

Solution:

(a) Each of the 100 selected people will be assigned a unique random number using a random number generator. A list of names and numbers will be created and soiled from smallest to largest by the assigned numbers (and carrying along the name). The *first 50* people on the list will be asked to apply the *new* compound to their right arm and the other 50 people will be asked to apply the *current* compound to their right arm. The compounds will be put in identical, unmarked tubes so neither the participants nor the researchers know which compound is being applied. The analysts will be the only people who know which participants received which compound. Each person will be randomly assigned to a bin by assigning random numbers to bins using a random number generator. The first person on the list will be assigned to the bin with the smallest number, the second person on the list will be assigned to the bin with the second smallest number, and so on. After each person inserts his or her right arm into the assigned bin for one minute, the number of mosquito bites will be counted.

The mean number of mosquito bites for the two compounds will be compared using a two-sample t-test and/or a confidence interval for the difference in means for two independent samples.

(b) Each participant will be randomly assigned to a bin as described in part (a). The researchers will distribute two identical tubes, one labeled 1 and the other labeled 2, to each participant. One of those tubes will contain the *new* compound and the other will contain the *current* compound. Neither the researchers nor the participants will know which compound is in which tube. Only the analyst will have this information. Each participant will apply one compound to one arm and the other compound to the other arm. The assignment of the compounds to the arms is completed using randomization. A random number will be generated for each participant, and the participants with the 50 smallest random numbers will apply tube 1 to their right arm, and the remaining 50 participants will apply tube 2 to their left arm. Each

participant will insert both arms into the assigned bin at the same time for one minute, and the number of mosquito bites will be counted on each arm.

The analyst will compute the difference in the number of bites (new - current) for each of the 100 participants and use a one-sample t-test and/or construct a confidence interval for the mean of the differences to test the null hypothesis that the mean difference is zero.

Comments. Notes for part (b):

Examples of alternative matched-pairs designs that will be scored as essentially correct:

1. If the student assumes that only one arm can be inserted into a bin, the experiment can be done by randomly selecting 50 of the 100 available participants. The bins are labeled from 1 to 100 and two bins are randomly assigned to each subject, one for each arm. This can be done by using a random number generator to assign each participant a unique random number and sorting the list of participants with respect to those numbers. Bins 1 and 2 are assigned to the first participant on the list, bins 3 and 4 are assigned to the second participant on the list, and so on. For each subject, a coin is tossed to determine which bin is assigned to which arm. Then, the experiment proceeds as described above. Another coin will be tossed to decide which arm receives the new compound. The current compound is applied to the other arm. Either both arms will be tested at the same time or another coin will be tossed to determine which arm is tested first.

this design, each participant uses one bin twice. One bin is randomly assigned to each of the 100 participants as described above. Each participant will apply one compound to one arm and the other compound to the other arm. A coin can be tossed to decide which arm receives the new compound. The current compound is applied to the other arm. The coin can be tossed a second time to determine which arm is put into the bin first. A potential disadvantage of this experiment is that the mosquitoes that are most aggressive and most resistant to the compounds will be more likely to bite the first arm inserted into a bin and less likely to bite the second arm. Randomizing the order in which the arms are inserted into the bin controls potential bias of this order effect, but the variability of the observed differences may be substantially larger than for the two other matched-pairs designs that were previously described, and this would make the comparison of the effects of the compounds less precise.

(c) The matched-pairs design in part (b) is better because one potential source of variation, person-to-person variability in susceptibility to mosquito bites, is controlled.

Problem 20. AP 2003 №4

Because of concerns about employee stress, a large company is conducting a study to compare two programs (tai chi or yoga) that may help employees reduce their stress

levels. Tai chi is a 1,200-year-old practice, originating in China, that consists of slow, fluid movements. Yoga is a practice, originating in India, that consists of breathing exercises and movements designed to stretch and relax muscles. The company has assembled a group of volunteer employees to participate in the study during the first half of their lunch hour each day for a 10-week period. Each volunteer will be assigned at random to one of the two programs. Volunteers will have their stress levels measured just before beginning the program and 10 weeks later at the completion of it.

(a) A group of volunteers who work together ask to be assigned to the same program so that they can participate in that program together. Give an example of a problem that might arise if this is permitted. Explain to this volunteer group why random assignment to the two programs will address this problem.

(b) Someone proposes that a control group be included in the design as well. The stress level would be measured for each volunteer assigned to the control group at the start of the study and again 10 weeks later. What additional information, if any, would this provide about the effectiveness of the two programs?

(c) Is it reasonable to generalize the findings of this study to all employees of this company? Explain.

Solution:

(a) For example, a deadline in the department where the group of volunteers works has been moved back, lowering the stress levels of those working in the department. If the volunteers from this department were all in the same treatment group, this *change in stress level* could mistakenly be attributed to the treatment.

Without random assignment of volunteers to the two programs, it is possible that the two treatment groups could differ in some way that affects the outcome of the experiment. Randomization "evens out" the possible effects of potentially confounding variables.

(b) Without the control group, the company could compare the two treatments, but would not be able to say whether the observed reduction in stress was attributable to participation in the programs. For example, a change in the work environment during this period might have reduced the stress level of all employees. The addition of a control group would enable the company to assess the magnitude of the mean reduction attributable to each treatment, as opposed to just determining if the two programs differ.

(c) It is not reasonable to generalize the findings of this study to all employees, because the participants in this experiment were volunteers and volunteers may not be representative of the population OR the participants were not randomly selected from the company employees.

Problem 21. AP 2003 Form B №4 a,b,d

There have been many studies recently concerning coffee drinking and cholesterol level. While it is known that several coffee-bean components can elevate blood cholesterol level, it is thought that a new type of paper coffee filter may reduce the presence of some of these components in coffee.

The effect of the new filter on cholesterol level will be stadied over a 10 week period using 300 non-smokers who each drink 4 cups of caffeinated coffee per day. Each of these 300 participants will be assigned to one of two groups experimental group, who will only drink coffee that has been made with the new filter, or the control group, who will only drink coffee that has been made with the standard filter. Each participant's cholesterol level will be measured at the beginning and at the end of the study.

(a) Describe an appropriate method for assigning the subject to the two groups so that each group will have an equal number of subjects.

(b) In this study, the researchers chose to include a group who only drank coffee that was made with the standard filter. Why is it important to include a control group in this study even though cholesterol levels will be measured at the beginning and at the end of the study?

(d) Why would the researchers choose to use only non-smokers in the study?

Solution:

(a) Assign each subject a number from 001 to 300.

Generate 150 random integers, ignoring repeats and out of order.

The subjects with those integers will go for the new filter group.

The remaining 150 would be assigned to the standard filter group.

(b) Without control group, the cholesterol level may change, but we will not be able to say whether it has changed by its own (by some other extraneous variables) or by the treatment during the 10-weeks period. For instance a diet of a person might change with time of the year, and the diet may result in change of cholesterol level. Such a change will not be reasoned by the new filter. Thus the addition of a control group enables the researcher to assess the change in cholesterol level due to the coffee filter alone, as opposed to overall change in cholesterol level.

(d) This eliminates smoking as a source of variability that may affect the change in cholesterol level. It will then be a solution to control for smoking using only nonsmokers. Doing this creates more homogeneous groups for more precise estimates of the treatment effects. The drawback of such a solution is that the researcher will be able to generalize results only to nonsmokers.

Practice AP problems

Problem 1. AP 2016 №3
Alzheimer's disease results in a loss of cognitive ability beyond what is expected with typical aging. A local newspaper published an article with the following headline.

Study finds strong association between smoking and Alzheimer's

The article reported that a study tracked the medical histories of 21,123 men and women for 23 years. The article stated that, for those who smoked at least two packs of cigarettes a day, the risk of developing Alzheimer's was 2.57 times the risk for those who did not smoke.

(a) Identify the explanatory and response variables in the study.

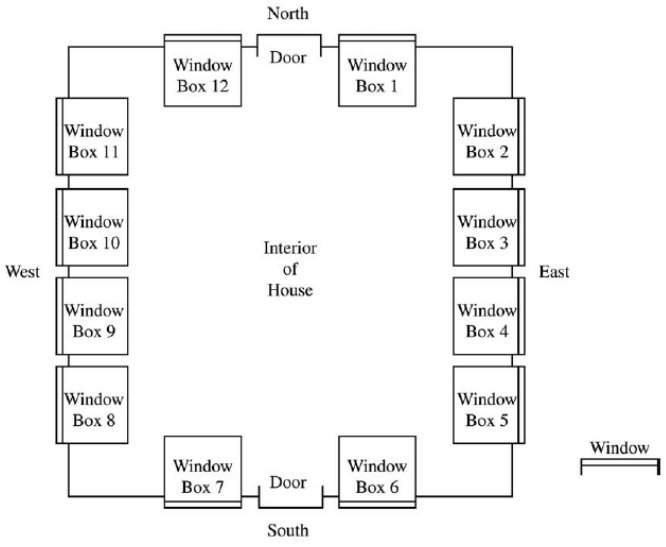
Explanatory variable:

Response variable:

(b) Is the study described in the article an observational study or an experiment? Explain.

(c) Exercise status (regular weekly exercise versus no regular weekly exercise) was mentioned in the article as a possible confounding variable. Explain how exercise status could be a confounding variable in this study.

Problrm 2. AP 2007 Form B №3
The United States Department of Energy is conducting an experiment to compare the heat gain in houses using two different types of windows, A and B. Six windows of each type are available for the experiment. The Department has constructed a house with twelve widows as shown on the floor plan below.



In the interior of the house, each window is surrounded by a window box to capture and measure the amount of heat coming in through that window and to isolate the heat of gain for each window.

(a) A randomized block experiment will be used to compare the heat gain for the two types (A and B) of windows. How would you group the window boxes into blocks? (Clearly indicate your blocks using the window box numbers.) Justify your choice of blocks.

(b) For the design in part (a), describe how you would assign window types (A and B) to the numbered window boxes.

Problem 3. AP 2005 №5 a,c A survey will be conducted to examine the educational level of adult heads of households in the United States. Each respondent in the survey will be placed into one of the following two categories:

- Does not have a high school diploma
- Has a high school diploma

The survey will be conducted using a telephone interview. Random-digit dialing will be used to select the sample.

(a) For this survey, state one potential source of bias and describe how it might affect the estimate of the proportion of adult heads of households in the United States who do not have a high school diploma.

(b) Topic: Confidence intervals

(c) Since education is largely the responsibility of each state, the agency wants to be sure that estimates are available for each state as well as for the nation. Identify a sampling method that will achieve this additional goal and briefly describe a way to select the survey sample using this method.

Problem 4. AP 2004 №2
Researchers who are studying a new shampoo formula plan to compare the condition of hair for people who use the new formula with the condition of hair for people who use the current formula. Twelve volunteers are available to participate in this study. Information on these volunteers (numbered 1 to 12) is shown in the table below.

Volunteer	Gender	Age
1	Male	21
2	Female	20
3	Male	47
4	Female	60
5	Female	62
6	Male	61
7	Male	58
8	Female	44
9	Male	44
10	Female	24
11	Male	23
12	Female	46

1. These researchers want to conduct an experiment involving the two formulas (new and current) of shampoo. They believe that the condition of hair changes with age but not gender. Because researchers want the size of the blocks in an experiment to be equal to the number of treatments, they will use blocks of size 2 in their experiment. Identify the volunteers (by number) that would be included in each of the six blocks and give the criteria you used to form the blocks.

2. Other researchers believe that hair condition differs with both age and gender. These researchers will also use blogs of size 2 in their experiment. Identify the volunteers (by number) that would be included in each of the six blocks and give the criteria you used to form the blocks.

3. The researchers in part (b) decide to select three of the six blocks to receive the new formula and to give the other three blocks the current formula. Is this an appropriate way to assign treatments? If so, describe a method for selecting the three blocks to receive the new formula. If not, describe an appropriate method for assigning treatments.

Problem 5. AP 2004 Form B №2

At a certain university, students who live in dormitories eat at a common dining hall. Recently, some students have been complaining about the quality of the food served there. The dining hall manager decided to do a survey to estimate the proportion of students living in the dormitories who think that the quality of the food should be improved. One evening, the manager asked the first 100 students entering the dining hall to answer the following question.

> Many students believe that the food served in the dining hall needs improvement. Do you think that the quality of food served here needs improvement, even though that would increase the cost of the meal plan?
> Yes No No opinion

(a) In this setting, explain how bias may have been introduced based on the way this convenience sample was selected and suggest how the sample could have been selected differently to avoid that bias.

(b) In this setting, explain how bias may have been introduced based on the way the question was worded and suggest how it could been worded differently to avoid that bias.

Problem 6. AP 2003 Form B №3 a

A study was conducted to determine if taking vitamin C reduces the occurrence of the flu. The study was conducted using 808 student volunteers who did not take a flu shot. The subjects were randomly assigned to one of two groups: a treatment group who received 1000 mg of vitamin C daily or a control group who received a placebo flavored to taste like the vitamin C treatment. All participants were monitored to

ensure that they adhered to their assigned treatment on a daily basis throughout the period of the study. At the end of the flu season, each subject's medical record was reviewed by a physician to determine whether he or she had contracted the flu during the period of the study. The physician did not know which treatment each subject received.

(a) Is this study an experiment or an observational study explain your answer

Answers to practice problems

Problem 1.

(a) Person's degree of smoking, whether the person develops Alzheimer's during study.

(b) Observational study due to study design. No assignment to treatment.

(c) Exercise is a lurking variable. Connection: exercise and will to smoke, exercise and likeliness of Alzheimer's. No connection: smoke and Alzheimer's.

Problem 2.

(a) Blocks 1 and 12; 2, 3, 4 and 5; 6 and 7; 8,9,10 and 11 OR 1-12, 2-3, 4-5, 6-7, 8-9, 10-11 with respect to exposure (side of house).

(b) Random assignment to treatment using flipping of coin within each block. treatment: window type A or B, experimental units: window boxes.

Problem 3.

1. The difference is provided among respondents and non-respondents in having diploma linked to whether or not a person has a phone. The estimated proportion might be under-estimated.

3. Stratified random sampling. State is a stratum.

Problem 4.

(a) 1-2, 10-11, 8-9, 3-12, 4-7, 5-6. Age.

(b) 2-10, 8-12, 4-5, 1-11, 3-9, 6-7. Gender, age.

(c) No. Random assignment to treatment within each block. Treatment: type of formula.

Problem 5.

1. Convenience sample. Respondents may differ with respect to opinion of food quality from nonrespondents. Respondents: students who arrived. Nonrespondents: students who did not arrive. SRS of 100 dormitory residents.

2. Non-neutral question: includes statement about opinion of majority. Fear of consequence: cost increase. "Do you think the quality of food served needs improvement?"

Problem 6.

(a) An experiment. There was random assignment to treatment. Treatmnet: vitamin C.