Unit 23
Experimental Design

Objectives:

- To understand the difference between an observational study and an experiment
- To understand how a completely randomized design is employed
- To understand how a randomized block design is employed
- To understand how to identify response variables and explanatory variables
- To understand how to identify extraneous variables and sources of bias which may influence the results in a study
- To understand the importance and the role of an institutional review board, informed consent, and confidentiality in a study involving human subjects

The hypothesis tests we have discussed thus far are based on the assumption that simple random sampling is used to collect data. We have noted though that for the purpose of statistical analysis, a sample may be treated as a simple random sample even though strictly speaking it may not be. For example, one might find it more convenient to use systematic random sampling as a substitute for simple random sampling, while expecting the systematic random sample to be at least as good as a simple random sample at representing the desired population.

However, in some situations, some type of a non-random may be the only option. For instance, if the purpose of a study is to draw some conclusions about the population of college students, it is very difficult, if not impossible, to obtain a truly simple random sample from this population. The large number of colleges spread over a wide geographical area make it unrealistic to attempt selecting a sample where every single college student has an equal chance of being in the sample. Even if this were possible, there is no guarantee that selected students will agree to participate in the study, and no one can ethically force individuals to agree to participate. Consequently, some sort of convenience sample is often used in such a study. The fact that a statistical analysis which is designed for use with a simple random sample is applied to a non-random sample has to be acknowledged as one of the limitations of the study. Every study, no matter how carefully done or how many resources are available, will generally have some limitations.

The items on which one or more variables are measured are called sampling units; sampling units which are human beings are called *subjects*. In some situations, data can be collected immediately upon obtaining one or more samples; in other situations, however, some intervention by a researcher must be employed before data can be collected. For instance, if we are comparing body temperature for males and females, then we may proceed to record body temperatures as soon as we have obtained our subjects; however, if we are comparing the time to perform a specific task with and without background music, we will not be able to record the time it takes for each of our subjects to perform the task until we first make some decisions: (1) Do we have each subject perform the task only once, some with background music and some without, or do we have each subject perform the task twice, once with background music and once without? (2) If we have each subject perform the task only once, which subjects will have background music and which will not; if we have each subject perform the task twice, will background music be supplied first or second?

An observational study is performed when data is collected by merely making observations with no intervention from a researcher. Comparing body temperature for males and females after observing and recording body temperatures from our subjects is an example of an observational study, since there is no intervention by a researcher when data is collected. An *experiment* is performed when data is collected after some intervention by the researcher is employed. Comparing the time to perform a specific task with and without background music after deciding whether subjects will be assigned to two separate groups or will each perform the task twice is an example of an experiment. The decisions concerning what interventions a researcher will perform before collecting data is called an *experimental design*.

We shall only discuss two very simple and basic experimental designs. To motivate the discussion about experimental design, we shall consider hypothesis tests where two or more population means are being compared. Our discussion of hypothesis testing concerning means thus far has involved only tests where one population mean is being compared to a hypothesized value. In many practical situations, however, it is of interest to compare two or means with each other.

Let us imagine that we want to see if there is any evidence that the mean time to perform a specific task is reduced when background music is provided. Before collecting data, we would of course have to obtain subjects for our study. As we have already mentioned, data can not be collected immediately upon obtaining subjects. We must first decide what type of experimental design will be used to collect data from the subjects. One way to obtain data for a hypothesis test would be to divide the available subjects randomly into two groups, where one group would perform the task with background music, and the other group would perform the task without background music. An alternative way to obtain data for a hypothesis test would be to allow each of the available subjects to perform the task twice, once with background music provided and once without background music, in random order. These two alternative methods to collect data are examples of two different types of experimental design.

A *completely randomized design* is one where sampling units (subjects) are randomly and independently assigned to two or more groups. A *randomized block design* is one where random assignment to two or more groups is done within *blocks*. Dividing the available subjects randomly into two groups, where one group performs the task with background music and the other group performs the task without background music is a completely randomized design. Allowing each of the available subjects to perform the task twice, once with background music provided and once without background music, in random order, is a randomized block design; the subjects are the blocks within which the randomization is done. These two different designs result in two different kinds of data.

Data from the completely randomized design consists of two independent samples of observations of times to complete the task. Data from the randomized block design consists of one sample of paired observations of times to complete the task. The distinction between these two approaches to comparing mean performance time with and without background music is that with two independent samples, our focus is on the difference between two sample means, but with one sample of paired data, our focus is on the mean of the differences between paired observations. Before any further discussion of the appropriate data analysis with each these two approaches, let us first consider exactly how the randomization can be done in each case. Suppose 14 subjects are available for the study to decide if there is any evidence that the mean time to perform a specific task is reduced when background music is provided.

To utilize a completely randomized design, we can make use of a random number table (or any other source of random number generation) to assign subjects to one group or the other by first labeling each of the 14 subjects with the integers

01, 02, 03, 04, 05, 06, 07, 08, 09, 10, 11, 12, 13, 14.

We then read seven two-digit entries from the random number table to obtain one group of subjects, and the remaining subjects would comprise the other group. As is always the case in reading a random number table, we skip repeats. If we were to read the first two digits of each set of five starting in the 21st row of the first page of Table A.1, the subjects selected for one group will be those labeled

12, 02, 05, 04, 14, 09, 07.

These subjects would then be assigned to one group, and the remaining subjects would be assigned to the other group.

To utilize a randomized block design, we can make use of a random number table to decide the order of the conditions under which each subject will perform the task twice. If we first label the subjects from 1 to 14, we could then read 14 digits from the random number table, one for each subject; each subject for whom an even digit was read from the table could be assigned to perform the task without background first and with background music second, and each subject for whom an odd digit was read from the table could be assigned to perform the task with background first and without background music second. If we were to read the first 14 digits in the 25th row of the first page of Table A.1, we would find

2, 3, 8, 9, 7, 1, 6, 9, 2, 8, 4, 2, 3, 8.

This would tell us that subject #1 will perform the task without background first and with background music second, subject #2 will perform the task with background first and without background music second, subject #3 will perform the task without background first and with background music second, etc. It is important that the order be randomized, so that if a significant difference is found, this difference can be attributed to whether or not background music was provided and not to the order in which background music was provided and not provided.

Recall that when we discussed linear regression, we defined a response variable to be a variable we predict from one or more other variables, and we defined an explanatory variable to be a variable from which we make predictions. We may generalize these definitions to include experimental design situations. We now define a *response variable* to be a variable whose changes are being studied, and we define an *explanatory variable* to be a variable which potentially influences changes in a response variable. An experimental design enables us to study the influence one or more explanatory variables exert on a response variable. In the study to see if there is any evidence that the mean time to perform a specific task is reduced when background music is provided, the response variable is "time to perform a the task," and the explanatory variable is "whether or not background music is present"; note that the response variable can be treated as quantitative, and the explanatory can be treated as qualitative-dichotomous.

It is impossible to identify all the variables which can potentially influence changes in a given response variable. An *extraneous variable* is one which we do not or can not measure directly but may potentially influence the response variable being studied; this can potentially lead to incorrect conclusions. For example, with the randomized block design where each of the available subjects to perform a task twice, once with background music provided and once without background music, the order in which background music was provided and not provided is an extraneous variable which has the potential to influence our results. Randomizing the order is an effort to reduce or eliminate the influence of this particular extraneous variable.

In trying to decide if mean time to perform a specific task is reduced when background music is provided, extraneous variables related to a subject's ability have the potential to influence our results. Will the completely randomized design or the randomized block design do a better job of eliminating the influence of such extraneous variables? To answer this question, first consider what might happen if two independent samples were used. It is possible that one sample contains a substantial number of individuals who are quicker at performing the task than those in the other sample. If the quicker individuals are assigned to the sample where subjects perform the task with background music provided, we might attribute a significant reduction in mean performance time to the background music, when in actual fact the background music may have had little or no effect; if the quicker individuals are assigned to the sample where subjects perform the task with no background music provided, we might fail to find a significant reduction which should be attributed to the background music. The random assignment of subjects to the two samples will tend to reduce but not eliminate the influence of such extraneous variables; however, allowing each subject to perform the task twice, once under each condition in random order, guarantees that the extraneous variables related to a subject's ability will be eliminated from influencing our results. We see then that, in general, a randomized block design is preferable to a completely randomized design, when possible.

There are situations where a randomized block design may not be possible though. Suppose we are comparing the lifetime of two brands of light bulbs, for instance. Lifetime is the response variable, and brand of light bulb is the explanatory variable; each bulb is a sampling unit. We certainly cannot assign a bulb to one brand for the purpose of measuring its lifetime, and afterward assign the same bulb to another brand to measure its lifetime again! Instead of paired data from a randomized block design, we must settle for data from two independent samples. That is, we obtain a simple random sample of bulbs from one brand and another simple random sample of bulbs from the other brand.

Even when we are unable to assign individual sampling units to one group or another, there may still be ways in which a randomized block design can be employed to eliminate the influence of an extraneous variable. Matching two or more sampling units which can be treated as similar (or identical) is one way to employ a randomized block design. For instance, let us imagine that we want to compare mean height for two types of plants. Height is the response variable, and type of plant is the explanatory variable; each seed (plant) is a sampling unit. We certainly cannot assign a seed to one type of plant for the purpose of measuring the height to which it grows, and then grow the same seed again as a different type of plant to measure its height again! However, we can design the experiment in a way which will generate paired data by eliminating the influence of extraneous variables related to the care each plant gets as it grows. By growing two plants, one of each type, in each of say $n = 20$ pots, we can measure two plant heights from each pot, yielding a sample of 20 paired measurements. The pots would be the blocks in this randomized block design. It is reasonable to assume that the two plants in any one pot would be given practically identical treatment in all respects, thus eliminating the influence of several extraneous variables on our results. Consequently, this randomized block design would be better than simply generating two independent samples of observations by growing one plant in each of 20 pots where ten were of one type, and ten were of the other type. (In studies involving animals or human beings, researchers may attempt to employ a randomized block design through the use of twins. Of course, getting twins as subjects in an experiment may not always be so easy in practice.)

Self-Test Problem 23-1. A study is being done to compare right-hand grip strength and left-hand grip strength in 21-year-old, right-handed males. Eight 21-year-old, right-handed males are available as subjects. Grip strength is measured by having each subject use one hand to grip two bars separated by a spring, bringing the bars as close together as possible.

(a) Is this an observational study or an experiment? Why?
(b) Identify the response variable, the explanatory variable, and the sampling unit.
(c) Describe how a completely randomized design would be employed to obtain data.
(d) Describe how a randomized block design would be employed to obtain data.
(e) Which of the two designs described in parts (c) and (d) would be preferable, and why?
(f) For each of eight 21-year-old, right-handed males, the right-hand grip strength and the left-hand grip strength are recorded in pounds of force with the following results:

Right	86	87	83	93	71	76	91	100
Left	84	84	76	90	67	70	87	99

Should this data be treated as one sample of paired observations or as two independent samples of observations?

Self-Test Problem 23-2. For each situation described, (i) identify the response variable, (ii) identify the explanatory variable, (iii) identify the sampling unit, and (iv) decide whether the data should be treated as one sample of paired observations or as two independent samples of observations.

(a) The variables "Weekly TV Hours" and "Weekly Radio Hours" in the SURVEY DATA, displayed as Data Set 1-1 at the end of Unit 1, are being used to see if there is any evidence of a difference in the mean time spent watching television weekly and the mean time spent listening to the radio weekly for voters in a state.
(b) The variables "Weekly TV Hours" and "Weekly Radio Hours" in the SURVEY DATA, displayed as Data Set 1-1 at the end of Unit 1, are being used to see if there is any evidence of a difference between males and females in the mean time spent watching television weekly for voters in a state.
(c) In order to see if there is any evidence that mean breaking strength is higher for Deluxe brand rope than for Econ brand rope, the breaking strength in pounds of force is recorded for 12 pieces of Deluxe brand rope and 10 pieces of Econ brand rope with the following results:

Deluxe	183	197	179	177	176	187	171	199	182	197	197	187
Econ	162	165	180	185	189	176	159	164	186	164		

Self-Test Problem 23-3. Each of parts (a), (b), and (c) in Self Test Problem 23-2 describes a hypothesis testing situation. Indicate whether the hypothesis test will be one-sided or two-sided. (Do not try to perform any of the steps of a hypothesis test.)

Our illustrations of completely randomized designs and randomized block designs thus far have applied to situations where two means are being compared, but these designs can be utilized even when more than two means are being compared. Let us suppose that each of five physicians identifies four patients suffering from a specific ailment, and these 20 patients are made available for a study which will be designed to compare a placebo and three different treatments of the ailment labeled A, B, and C. A *placebo* is a dummy preparation having no medical value, but it is given to subjects who believe it is a genuine treatment. (It has been claimed that the psychological effects of believing that a treatment is genuine may produce very real and observable improvement in one's condition!) The condition of a subject is the response variable, and type of treatment is the explanatory variable; each patient selected is a subject (sampling unit).

Each subject will be treated with one of A, B, C, or the placebo. After one month, each subject will be examined by his or her physician. The physician will then assign to each subject examined a score reflecting how much change has occurred in the subject's condition, improvement or otherwise. The mean of the scores assigned to the subjects at the end of one month is being compared with regard to the type of medication administered.

Regardless of what type of design is used in this experiment, there could be a concern that knowledge of which treatment was administered to a subject could unconsciously influence a physician's scoring of that subject. For instance, one of the physicians may hold a personal belief that one treatment is better than another or may not expect any improvement in a subject who was given the placebo. If both the physicians and the subjects are given no information on what is administered, the experiment is called *double-blind*; if only the subjects are given no information on what is administered, the experiment is called *single-blind*. We next consider how we might perform this experiment using a completely randomized design and using a randomized block design.

We would employ a completely randomized design by randomly assigning five of the 20 subjects to each of the four medications. One way to accomplish this would be to label each subject with a double digit integer and assign the first 4 labels read from a random number table (or any other source of random number generation) to treatment A, the next 4 labels read to treatment B, the next 4 labels read to treatment C, and the remaining labels to the placebo.

Among the extraneous variables which can influence the subjects' scores is which physician does the examination. Even in a double-blind experiment, it is not inconceivable that there will be some differences among physicians in scoring. While a completely randomized design is likely to reduce the influence of this extraneous variable on the response variable (the scores), a randomized block design can guarantee that this extraneous variable will not have any influence.

We could employ a randomized block design by randomly assigning one of the four subjects identified by one particular physician to each of the four medications; then we repeat this procedure for each of the other five physicians. In other words, the physicians would be the blocks. With a random number table (or any other source of random number generation) we could easily make the assignments for each physician's 4 subjects by first labeling each of the four subjects with one of the integers 1, 2, 3, and 4. We then assign the first label read in the random number table (or other source of random number generation) to treatment A, the second label read to treatment B, the third label read to treatment C, and the remaining label to the placebo. This design guarantees that each physician will score one subject from each medication administered.

The randomized block design results in four observations per physician, one for each treatment. If there were only two observations per physician, we would have paired data, but when two or more measurements are made from the same source (such as the physicians in our illustration), the resulting data can be called *repeated measures* data. Paired data is the special case of repeated measures where only two repetitions are made on each source.

Many complex randomization schemes are possible depending on the number and types of variables involved. The type of experimental design used will affect the decision about what hypothesis test is appropriate. We shall be considering hypothesis tests only for completely randomized designs, which generate two or more independent samples of data, and for randomized block designs, which generate paired data.

Thus far, we have really only addressed technical aspects of experimental design. In a study involving human subjects, ethical issues must also be considered. Studies funded by the federal government are required by law to satisfy the following principles: (1) an *institutional review board* must review the plan for the study to protect subjects from possible harm; (2) all subjects in the study must give their *informed consent* prior to data collection; (3) individual data must be kept confidential, and only group summary statistics may be reported publicly. Let us consider these principles with regard to the earlier illustration where each of five physicians identifies four patients suffering from a specific ailment. There are 20 subjects (patients) made available for a study which is designed to compare a placebo and three different treatments of the ailment labeled A, B, and C.

An institutional review board will be concerned only with the manner by which data is to be collected from subjects, the information given to subjects about potential risks and benefits, and the manner by which a subject's consent is obtained; it is not the purpose of the institutional review board to consider the design of the study or the value of the knowledge to be gained from the study. It is federal policy to include at least one non-scientist as part of the board to guard against a desire for new knowledge reducing the inclination to criticize the plan for the study. In the study to compare a placebo and treatments A, B, and C, subjects need to be informed about the denial of a potentially valuable treatment or the administration of untested treatments before giving consent to be part of a study. Furthermore, consent should be obtained without coercion, and subjects should have the right to withdraw from the study at any time.

Ensuring that informed consent is obtained from subjects can be a complex process. Even when every effort is made to indicate that participation in a study is purely voluntary, potential subjects may still suspect that they will be penalized or denied some benefit(s), if they do not participate in the study. Whether or not a potential subject has the ability to give informed consent is an important consideration; for instance, a child or a mentally impaired person may not be capable of providing informed consent without some assistance. In order to document informed consent, subjects are generally asked to sign a consent form. It is important that the consent form include complete information that subjects need to know about the study. A researcher might on a consent form include a discussion about the value of the results of the study in order to encourage potential subjects to participate; however, the longer the consent form is, the less likely people will be to take the time to read it carefully. To entice potential subjects to participate in a study without having them feel pressured, some form of compensation might be offered. If the data to be collected is of a sensitive nature, the researcher may offer potential subjects anonymity in addition to confidentiality; this is easily accomplished by identifying each subject only by a code and separating names and other identifying characteristics from the data collected.

The difficulties with obtaining human subjects for a study will contribute to the limitations of the study. The population from which subjects are obtained (i.e., the accessible population) may not be exactly the same as the population to which the researcher would ideally want the results to apply. If the characteristics which make a subject less inclined to participate in a study are related to one or more of the variables being studied, the results of the study could be biased; a similar comment can be made concerning the characteristics which make a subject more inclined to withdraw from the study.

Another potential source of bias in a study comes from the fact that subjects' behavior could tend to be influenced merely by the fact that they are aware that they are part of a study; this is known as the *Hawthorne effect*. The Hawthorne effect was named after the Hawthorne Works of the Western Electric Company in the 1920s. When experiments were performed to study what changes in working conditions would improve productivity, it was found that any change made while the workers were aware they were part of a study (even negative changes) resulted in at least some increase in productivity! The amount of information to give to subjects in a study can be problematic. On the one hand, informed consent suggests that no vital information is withheld from subjects; on the other hand, specific information must be withheld from subjects in order to design a single-blind or double-blind experiment.

Some studies involve animal subjects instead of human subjects, but animals are certainly not capable of giving informed consent to be part of a study. Two issues concerned with the ethics of using animals for experimentation are the conditions under which animals should be used for experimentation and the safeguards which should be employed to ensure that animals are treated humanely in experiments.

Answers to Self-Test Problems

23-1 (a) This is an experiment, since data is collected with intervention from a researcher, the researcher having each subject use one hand to grip two bars separated by a spring. (b) "Right-hand grip strength" is the response variable, and "which hand is used" is the explanatory variable; each of the eight 21-year-old male subjects is a sampling unit. (c) The subjects would be divided randomly into two groups; the right-hand grip strength would be measured for each subject in one group, and the left-hand grip strength would be measured for each subject in the other group. (d) The right-hand grip strength and the left-hand grip strength would be measured (in random order) for each of the available subjects. (e) The randomized block design would be preferable, because the influence of extraneous variables on the results can be greatly reduced. (f) one sample of paired observations

23-2 (a) (i) response variable: "weekly hours" (ii) explanatory variable: "television or radio"
(iii) sampling units: the voters in the state (iv) one sample of paired observations
(b) (i) response variable: "weekly tv hours" (ii) explanatory variable: "sex"
(iii) sampling units: the voters in the state (iv) two independent samples of observations
(c) (i) response variable: "breaking strength" (ii) explanatory variable: "brand of rope"
(iii) sampling units: the pieces of rope (iv) two independent samples of observations

23-3 (a) two-sided (b) two-sided (c) one-sided

Summary

The items on which one or more variables are measured are called *sampling units*; sampling units which are human beings are called *subjects*. An *observational study* is performed when data is collected by merely making observations with no intervention from a researcher; an *experiment* is performed when data is collected after some intervention by the researcher is employed. The decisions concerning what interventions a researcher will perform before collecting data to reduce or eliminate the influence of extraneous variables is called an *experimental design*. A *completely randomized design* is one where sampling units (subjects) are randomly and independently assigned to two or more groups. A *randomized block design* is one where random assignment to two or more groups is done within *blocks*. Random numbers generated by a computer or programmable calculator can be used to make the necessary random assignments.

A *response variable* is a variable whose changes are being studied, and an *explanatory variable* is a variable which potentially influences changes in a response variable. An experimental design enables us to study the influence one or more explanatory variables exert on a response variable. An *extraneous variable* is one which we do not or can not measure directly but may potentially influence the response variable being studied. In general, a randomized block design is preferable to a completely randomized design, when possible, because a randomized block design guarantees that certain extraneous variables will not influence results, whereas a completely randomized design can only reduce the influence of these extraneous variables.

Completely randomized designs and randomized block designs can be utilized when two or more means are being compared. Data resulting from a completely randomized design can be treated as two or more independent simple random samples. Data resulting from a randomized block design can be called *repeated measures* data; paired data is the special case of repeated measures where only two repetitions are made on each source.

In experiments where subjects are assigned to treatments and human evaluators score the subjects, the experiment is called *double-blind* if both the evaluators and the subjects are given no information on what treatment is administered; if only the subjects are given no information on what treatment is administered, the experiment is called *single-blind*. Studies involving human subjects funded by the federal government are required by law to satisfy the following principles: (1) an *institutional review board* must review the plan for the study to protect subjects from possible harm; (2) all subjects in the study must give their *informed consent* prior to data collection; (3) individual data must be kept confidential, and only group summary statistics may be reported publicly. The difficulties with obtaining human subjects for a study can result in several limitations of the study: the population to which the results to apply, bias resulting from subjects who refuse to participate or

who drop out of the study, and bias resulting from the fact that subjects' behavior will tend to be influenced merely by the fact that they are aware that they are part of a study, known as the *Hawthorne effect*.

Animals are not capable of giving informed consent to be part of a study. Nevertheless, it is important to consider the conditions under which animals should be used for experimentation and the safeguards which should be employed to ensure that animals are treated humanely in experiments.

Data Set 23-1
CHAIN DATA

On the west coast of the United States is a chain of restaurants known as McDoogle's. Information about expenses, sales, and number of customers for the past year is gathered for a sample of restaurants in the northern part of the chain, and for a sample of restaurants in the southern part of the chain. For some restaurants, number of customers for the past year was not available.

(Expenses and Sales are in millions of dollars)

ID No.	Location	Expenses	Sales	Millions of Customers
01	South	1.0	0.1	3.5
02	North	1.2	1.8	4.4
03	South	2.8	4.0	
04	North	1.9	6.1	
05	South	0.3	5.3	4.2
06	South	1.5	4.0	0.5
07	South	3.4	7.4	5.1
08	North	1.0	3.6	3.7
09	South	1.6	2.2	
10	North	0.9	7.5	3.5
11	North	1.6	4.6	3.1
12	North	1.9	8.0	
13	North	0.6	3.3	3.8
14	North	1.0	2.5	3.5
15	South	2.5	2.6	2.4

ID No.	Location	Expenses	Sales	Millions of Customers
16	North	1.0	8.1	3.9
17	South	1.9	1.7	1.5
18	North	1.4	6.7	3.5
19	North	0.7	3.8	
20	North	1.0	5.2	4.4
21	North	1.3	7.8	
22	North	2.4	5.1	4.7
23	North	1.2	7.8	3.7
24	North	1.7	7.7	4.5
25	South	1.4	2.8	3.7
26	North	1.0	4.9	3.5
27	North	1.2	8.0	
28	South	2.3	2.0	2.9
29	North	2.2	2.5	4.1
30	North	0.8	5.1	4.2

Data Set 23-2
STUDENT DATA

For a sample of high school students, data on area of residence (rsd), hours spent watching TV per week (tvh), hours spent studying per week (std), and high school grade point average (gpa) are recorded.

Residence (rsd)	Weekly TV Hours (tvh)	Weekly Study Hours (std)	High School GPA (gpa)
urban	29	17	3.11
rural	21	27	3.25
urban	48	9	2.08
rural	30	22	3.31
rural	28	28	2.79
suburban	25	27	3.40
rural	11	46	3.95
urban	36	16	2.23
rural	14	32	2.84
suburban	39	14	2.57
suburban	41	2	2.46
urban	20	35	3.42
rural	13	41	3.36
suburban	38	17	2.94
urban	24	21	2.58
suburban	30	29	2.88
urban	23	26	2.60
suburban	34	21	2.25
urban	40	12	2.01
urban	33	19	2.53
rural	17	29	3.56
urban	44	10	1.97
rural	33	19	2.99
suburban	18	40	3.74
suburban	22	37	3.21
suburban	24	31	2.55
urban	27	21	3.20

Data Set 23-3
GPA DATA

For a sample of freshmen at a junior college, high school grade point average (hsgpa), high school class percentile rank (cpr), verbal SAT (vsat), math SAT (msat), and college freshman year grade point average (cgpa) are recorded.

High School GPA (hsgpa)	Class Percentile Rank (cpr)	Verbal SAT (vsat)	Math SAT (msat)	Freshman Year College GPA (cgpa)
3.38	80	650	530	3.26
4.23	89	610	660	3.48
3.95	95	630	520	3.59
2.89	62	460	530	1.84
3.26	96	540	640	2.77
3.25	55	590	690	2.22
3.92	77	640	600	3.53
3.79	91	590	530	3.22
3.33	66	610	610	2.82
3.20	73	700	690	3.46
3.70	71	560	620	2.53
3.43	57	490	500	1.98
3.38	60	490	600	2.70
3.40	78	710	580	3.48
3.12	61	510	510	2.59
2.71	60	560	490	2.04
2.92	55	610	510	2.58
3.66	79	550	520	2.90
2.63	58	660	570	2.50
2.87	59	620	600	2.46
3.72	60	520	630	2.44
3.18	83	520	610	2.61
2.83	70	600	630	2.82
3.35	92	630	560	3.23