

Maternal Smoking and Infant Health

WEDNESDAY, MARCH 1, 1995

***** New York Times

Infant Deaths Tied to Premature Births

Low weights not solely to blame

A new study of more than 7.5 million births has challenged the assumption that low birth weights per se are the cause of the high infant mortality rate in the United States. Rather, the new findings indicate, prematurity is the principal culprit.

Being born too soon, rather than too small, is the main underlying cause of stillbirth and infant deaths within four weeks of birth.

Each year in the United States about 31,000 fetuses die before delivery and 22,000 newborns die during the first 27 days of life.

The United States has a higher infant mortality rate than those in 19 other countries, and this poor standing has long been attributed mainly to the large number of babies born too small, including a large proportion who are born "small for date," or weighing less than they should for the length of time they were in the womb.

The researchers found that American-born babies, on

average, weigh less than babies born in Norway, even when the length of the pregnancy is the same. But for a given length of pregnancy, the lighter American babies are no more likely to die than are the slightly heavier Norwegian babies.

The researchers, directed by Dr. Allen Wilcox of the National Institute of Environmental Health Sciences in Research Triangle Park, N.C., concluded that improving the nation's infant mortality rate would depend on preventing preterm births, not on increasing the average weight of newborns.

Furthermore, he cited an earlier study in which he compared survival rates among low-birth-weight babies of women who smoked during pregnancy.

Ounce for ounce, he said, "the babies of smoking mothers had a higher survival rate". As he explained this paradoxical finding, although smoking interferes with weight gain, it does not shorten pregnancy.

Introduction

One of the U.S. Surgeon General's health warnings placed on the side panel of cigarette packages reads:

Smoking by pregnant women may result in fetal injury, premature birth, and low birth weight.

In this lab, you will have the opportunity to compare the birth weights of babies born to smokers and nonsmokers in order to determine whether they corroborate the Surgeon General's warning. The data provided here are part of the Child Health and Development Studies (CHDS)—a comprehensive investigation of all pregnancies that occurred between 1960 and 1967 among women in the Kaiser Foundation Health Plan in the San Francisco–East Bay area (Yerushalmy [Yer71]). This study is noted for its unexpected findings that ounce for ounce the babies of smokers did not have a higher death rate than the babies of nonsmokers.

Despite the warnings of the Surgeon General, the American Cancer Society, and health care practitioners, many pregnant women smoke. For example, the National Center for Health Statistics found that 15% of the women who gave birth in 1996 smoked during their pregnancy.

Epidemiological studies (e.g., Merkatz and Thompson [MT90]) indicate that smoking is responsible for a 150 to 250 gram reduction in birth weight and that smoking mothers are about twice as likely as nonsmoking mothers to have a low-birth-weight baby (under 2500 grams). Birth weight is a measure of the baby's maturity. Another measure of maturity is the baby's gestational age, or the time spent in the womb. Typically, smaller babies and babies born early have lower survival rates than larger babies who are born at term. For example, in the CHDS group, the rate at which babies died within the first 28 days after birth was 150 per thousand births for infants weighing under 2500 grams, as compared to 5 per thousand for babies weighing more than 2500 grams.

The Data

The data available for this lab are a subset of a much larger study — the Child Health and Development Studies (Yerushalmy [Yer64]). The entire CHDS database includes all pregnancies that occurred between 1960 and 1967 among women in the Kaiser Foundation Health Plan in Oakland, California. The Kaiser Health Plan is a prepaid medical care program. The women in the study were all those enrolled in the Kaiser Plan who had obtained prenatal care in the San Francisco–East Bay area and who delivered at any of the Kaiser hospitals in Northern California.

In describing the 15,000 families that participated in the study, Yerushalmy states ([Yer64]) that

The women seek medical care at Kaiser relatively early in pregnancy. Two-thirds report in the first trimester; nearly one-half when they are pregnant for

2 months or less. The study families represent a broad range in economic, social and educational characteristics. Nearly two-thirds are white, one-fifth negro, 3 to 4 percent oriental, and the remaining are members of other races and of mixed marriages. Some 30 percent of the husbands are in professional occupations. A large number are members of various unions. Nearly 10 percent are employed by the University of California at Berkeley in academic and administrative posts, and 20 percent are in government service. The educational level is somewhat higher than that of California as a whole, as is the average income. Thus, the study population is broadly based and is not atypical of an employed population. It is deficient in the indigent and the very affluent segments of the population since these groups are not likely to be represented in a prepaid medical program.

At birth, measurements on the baby were recorded. They included the baby's length, weight, and head circumference. Provided here is a subset of this information collected for 1236 babies — those baby boys born during one year of the study who lived at least 28 days and who were single births (i.e., not one of a twin or triplet). The information available for each baby is birth weight and whether or not the mother smoked during her pregnancy. These variables and sample observations are provided in Table 1.1.

Background

Fetal Development

The typical gestation period for a baby is 40 weeks. Those born earlier than 37 weeks are considered preterm. Few babies are allowed to remain in utero for more than 42 weeks because brain damage may occur due to deterioration of the placenta. The placenta is a special organ that develops during pregnancy. It lines the wall of the uterus, and the fetus is attached to the placenta by its umbilical cord (Figure 1.1). The umbilical cord contains blood vessels that nourish the fetus and remove its waste.

TABLE 1.1. Sample observations and data description for the 1236 babies in the Child Health and Development Studies subset.

Birth weight	120	113	128	123	108	136	138	132
Smoking status	0	0	1	0	1	0	0	0

Variable	Description
Birth weight	Baby's weight at birth in ounces. (0.035 ounces = 1 gram)
Smoking status	Indicator for whether the mother smoked (1) or not (0) during her pregnancy.

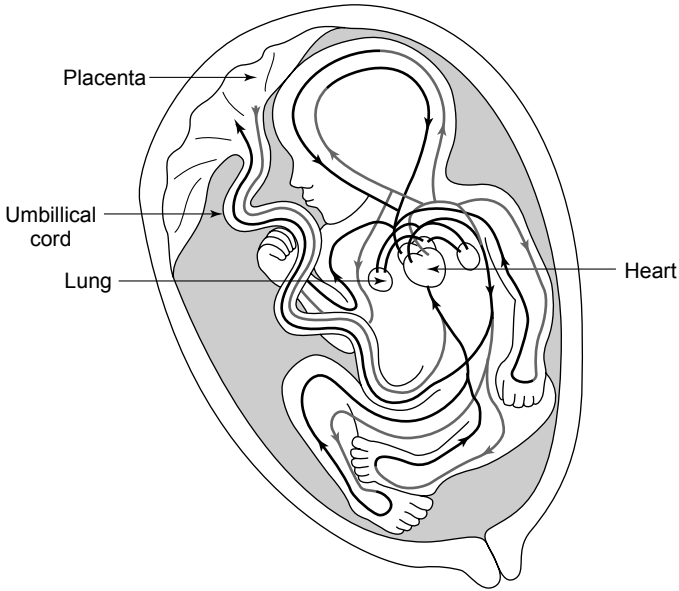


FIGURE 1.1. Fetus and placenta.

At 28 weeks of age, the fetus weighs about 4 to 5 pounds (1800 to 2300 grams) and is about 40 centimeters (cm) long. At 32 weeks, it typically weighs 5 to 5.5 pounds (2300 to 2500 grams) and is about 45 cm long. In the final weeks prior to delivery, babies gain about 0.2 pounds (90 grams) a week. Most newborns range from 45 to 55 cm in length and from 5.5 to 8.8 pounds (2500 to 4000 grams). Babies born at term that weigh under 5.5 pounds are considered small for their gestational age.

Rubella

Before the 1940s, it was widely believed that the baby was in a protected state while in the uterus, and any disease the mother contracted or any chemical that she used would not be transmitted to the fetus. This theory was attacked in 1941 when Dr. Norman Gregg, an Australian ophthalmologist, observed an unusually large number of infants with congenital cataracts. Gregg checked the medical history of the mothers' pregnancies and found that all of them had contracted rubella in the first or second month of their pregnancy. (There had been a widespread and severe rubella epidemic in 1940.) In a presentation of his findings to the Ophthalmological Society of Australia, Gregg ([Gre41]) replied to comments on his work saying that

... he did not want to be dogmatic by claiming that it had been established the cataracts were due solely to the "German measles." However, the evidence afforded by the cases under review was so striking that he was convinced that

there was a very close relationship between the two conditions, particularly because in the very large majority of cases the pregnancy had been normal except for the “German measles” infection. He considered that it was quite likely that similar cases may have been missed in previous years either from casual history-taking or from failure to ascribe any importance to an exanthem [skin eruption] affecting the mother so early in her pregnancy.

Gregg was quite right. Oliver Lancaster, an Australian medical statistician, checked census records and found a concordance between rubella epidemics and later increase in registration at schools for the deaf. Further, Swan, a pediatrician in Australia, undertook a series of epidemiological studies on the subject and found a connection between babies born to mothers who contracted rubella during the epidemic while in their first trimester of pregnancy and heart, eye, and ear defects in the infant.

A Physical Model

There are many chemical agents in cigarette smoke. We focus on one: carbon monoxide. It is commonly thought that the carbon monoxide in cigarette smoke reduces the oxygen supplied to the fetus. When a cigarette is smoked, the carbon monoxide in the inhaled smoke binds with the hemoglobin in the blood to form carboxyhemoglobin. Hemoglobin has a much greater affinity for carbon monoxide than oxygen. Increased levels of carboxyhemoglobin restrict the amount of oxygen that can be carried by the blood and decrease the partial pressure of oxygen in blood flowing out of the lungs. For the fetus, the normal partial pressure in the blood is only 20 to 30 percent that of an adult. This is because the oxygen supplied to the fetus from the mother must first pass through the placenta to be taken up by the fetus’ blood. Each transfer reduces the pressure, which decreases the oxygen supply.

The physiological effects of a decreased oxygen supply on fetal development are not completely understood. Medical research into the effect of smoking on fetal lambs (Longo [Lon76]) provides insight into the problem. This research has shown that slight decreases in the oxygen supply to the fetus result in severe oxygen deficiency in the fetus’ vital tissues.

A steady supply of oxygen is critical for the developing baby. It is hypothesized that, to compensate for the decreased supply of oxygen, the placenta increases in surface area and number of blood vessels; the fetus increases the level of hemoglobin in its blood; and it redistributes the blood flow to favor its vital parts. These same survival mechanisms are observed in high-altitude pregnancies, where the air contains less oxygen than at sea level. The placenta at high altitude is larger in diameter and thinner than a placenta at sea level. This difference is thought to explain the greater frequency in high-altitude pregnancies of abruptia placenta, where the placenta breaks away from the uterine wall, resulting in preterm delivery and fetal death (Meyer and Tonascia [MT77]).

Is the Difference Important?

If a difference is found between the birth weights of babies born to smokers and those born to nonsmokers, the question of whether the difference is important to the health and development of the babies needs to be addressed.

Four different death rates — fetal, neonatal, perinatal, and infant — are used by researchers in investigating babies’ health and development. Each rate refers to a different period in a baby’s life. The first is the fetal stage. It is the time before birth, and “fetal death” refers to babies who die at birth or before they are born. The term “neonatal” denotes the first 28 days after birth, and “perinatal” is used for the combined fetal and neonatal periods. Finally, the term “infant” refers to a baby’s first year, including the first 28 days from birth.

In analyzing the pregnancy outcomes from the CHDS, Yerushalmy ([Yer71]) found that although low birth weight is associated with an increase in the number of babies who die shortly after birth, the babies of smokers tended to have much lower death rates than the babies of nonsmokers. His calculations appear in Table 1.2. Rather than compare the overall mortality rate of babies born to smokers against the rate for babies born to nonsmokers, he made comparisons for smaller groups of babies. The babies were grouped according to their birth weight; then, within each group, the numbers of babies that died in the first 28 days after birth for smokers and nonsmokers were compared. To accommodate the different numbers of babies in the groups, rates instead of counts are used in making the comparisons.

The rates in Table 1.2 are not adjusted for the mother’s age and other factors that could potentially misrepresent the results. That is, if the mothers who smoke tend to be younger than those who do not smoke, then the comparison could be unfair to the nonsmokers because older women, whether they smoke or not, have more problems in pregnancy. However, the results agree with those from a Missouri study (see the left plot in Figure 1.2), which did adjust for many of these factors (Malloy et al. [MKLS88]). Also, an Ontario study (Meyer and Tonascia [MT77]) corroborates the CHDS results. This study found that the risk of neonatal death for babies who were born at 32+ weeks gestation is roughly the same for smokers and

TABLE 1.2. Neonatal mortality rates per 1000 births by birth weight (grams) for live-born infants of white mothers, according to smoking status (Yerushalmy [Yer71]).

Weight category	Nonsmoker	Smoker
≤ 1500	792	565
1500–2000	406	346
2000–2500	78	27
2500–3000	11.6	6.1
3000–3500	2.2	4.5
3500+	3.8	2.6

Note: 1500 to 2000 grams is roughly 53 to 71 ounces.

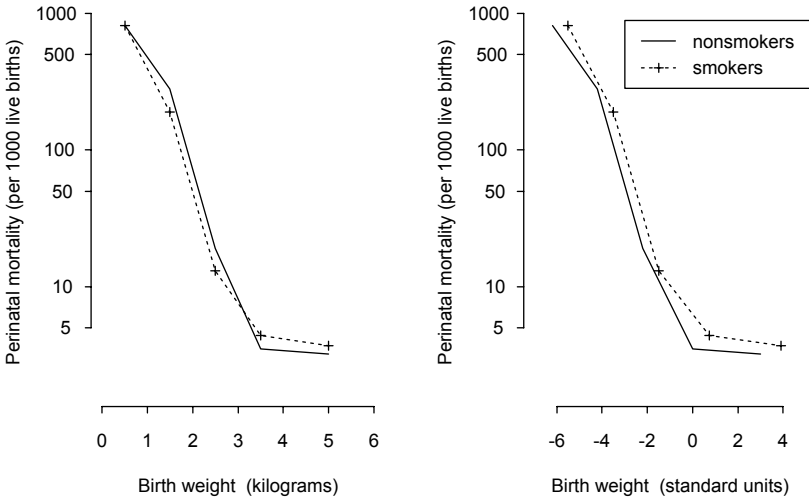


FIGURE 1.2. Mortality curves for smokers and nonsmokers by kilograms (left plot) and by standard units (right plot) of birth weight for the Missouri study (Wilcox [Wil93]).

nonsmokers. It was also found that the smokers had a higher rate of very premature deliveries (20–32 weeks gestation), and so a higher rate of early fetal death.

As in the comparison of Norwegian and American babies (New York Times, Mar. 1, 1995), in order to compare the mortality rates of babies born to smokers and those born to nonsmokers, Wilcox and Russell ([WR86]) and Wilcox ([Wil93]) advocate grouping babies according to their relative birth weights. A baby's relative birth weight is the difference between its birth weight and the average birth weight for its group as measured in standard deviations (SDs); it is also called the standardized birth weight. For a baby born to a smoker, we would subtract from its weight the average birth weight of babies born to smokers (3180 grams) and divide this difference by 500 grams, the SD for babies born to smokers. Similarly, for babies born to nonsmokers, we standardize the birth weights using the average and SD for their group, 3500 grams and 500 grams, respectively. Then, for example, the mortality rate of babies born to smokers who weigh 2680 grams is compared to the rate for babies born to nonsmokers who weigh 3000 grams, because these weights are both 1 SD below their respective averages. The right plot in Figure 1.2 displays in standard units the mortality rates from the left plot. Because the babies born to smokers tend to be smaller, the mortality curve is shifted to the right relative to the nonsmokers' curve. If the babies born to smokers are smaller but otherwise as healthy as babies born to nonsmokers, then the two curves in standard units should roughly coincide. Wilcox and Russell found instead that the mortality curve for smokers was higher than that for nonsmokers; that is, for babies born at term, smokers have higher rates of perinatal mortality in every standard unit category.

Investigations

What is the difference in weight between babies born to mothers who smoked during pregnancy and those who did not? Is this difference important to the health of the baby?

- Summarize numerically the two distributions of birth weight for babies born to women who smoked during their pregnancy and for babies born to women who did not smoke during their pregnancy.
- Use graphical methods to compare the two distributions of birth weight. If you make separate plots for smokers and nonsmokers, be sure to scale the axes identically for both graphs.
- Compare the frequency, or incidence, of low-birth-weight babies for the two groups. How reliable do you think your estimates are? That is, how would the incidence of low birth weight change if a few more or fewer babies were classified as low birth weight?
- Assess the importance of the differences you found in your three types of comparisons (numerical, graphical, incidence).

Summarize your investigations for the CHDS babies. Include the most relevant graphical output from your analysis. Relate your findings to those from other studies.

Theory

In this section, several kinds of summary statistics are briefly described. When analyzing a set of data, simple summaries of the list of numbers can bring insight about the data. For example, the mean and the standard deviation are frequently used as numerical summaries for the location and spread of the data. A graphical summary such as a histogram often provides information on the shape of the data distribution, such as symmetry, modality, and the size of tails.

We illustrate these statistics with data from the 1236 families selected for this lab from the Child Health and Development Study (CHDS). The data used here are described in detail in the Data section of the continuation of this lab in Chapter 10. For each statistic presented, any missing data are ignored, and the number of families responding is reported.

The Histogram

Figure 1.3 displays a histogram for the heights of mothers in the CHDS. The histogram is unimodal and symmetric. That is, the distribution has one mode (peak), around 64 inches, and the shape of the histogram to the left of the peak looks roughly like the mirror image of the part of the histogram to the right of the

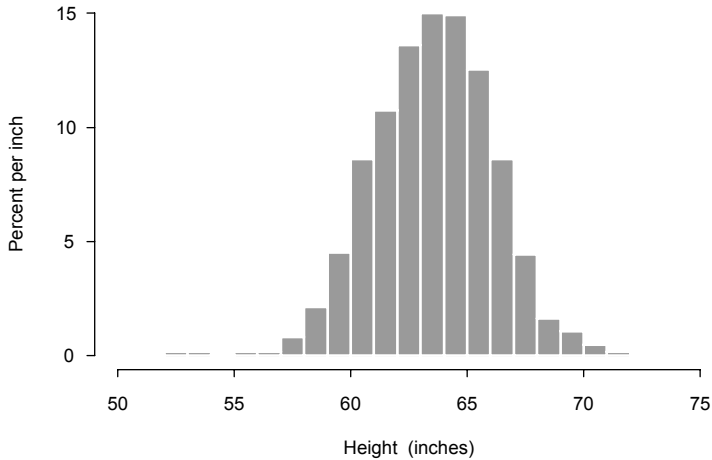


FIGURE 1.3. Histogram of mother's height for 1214 mothers in the CHDS subset.

peak. Outliers can be detected via histograms as well. They are observations that fall well outside the main range of the data. There appear to be a few very short mothers in the study.

In contrast to the height distribution, the histogram of the number of cigarettes smoked per day for those mothers who smoked during their pregnancy has a very different appearance (Figure 1.4). It shows two modes, one at 5–10 cigarettes and the other at 20–30 cigarettes. The distribution is asymmetric; that is it is right-skewed with the mode around 20–30 cigarettes less peaked than the mode at 0–5 cigarettes and with a long right tail. For unimodal histograms, a right-skewed distribution has more area to the right of the mode in comparison with that to the left; a left-skewed distribution has more area to the left.

A histogram is a graphical representation of a distribution table. For example, Table 1.3 is a distribution table for the number of cigarettes smoked a day by mothers who smoked during their pregnancy. The intervals include the left endpoint but not the right endpoint; for example the first interval contains those mothers who smoke up to but not including 5 cigarettes a day. In the histogram in Figure 1.4, the area of each bar is proportional to the percentage (or count) of mothers in the corresponding interval. This means that the vertical scale is percent per unit of measurement (or count per unit). The bar over the interval from 0 to 5 cigarettes is 3.2% per cigarette in height and 5 cigarettes in width: it includes all women who reported smoking up to an average of 5 cigarettes a day. Hence the area of the bar is

$$5 \text{ cigarettes} \times 3.2\%/\text{cigarette} = 16\%.$$

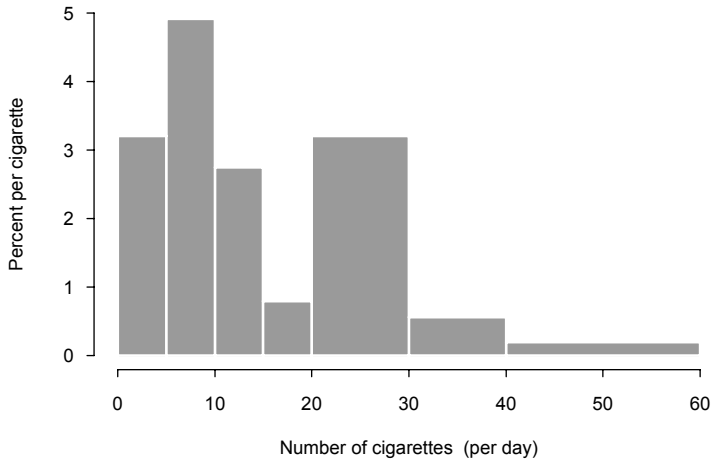


FIGURE 1.4. Histogram of the number of cigarettes smoked per day for the 484 mothers who smoked in the CHDS subset.

TABLE 1.3. Distribution of the number of cigarettes smoked per day for 484 mothers in the CHDS subset who smoked during their pregnancy, rounded to the nearest percent.

Number of cigarettes	Percent of smokers
0-5	16
5-10	25
10-15	14
15-20	4
20-30	32
30-40	5
40-60	4
Total	100

This bar is the same height as the bar above 20–30 cigarettes even though it has twice the number of mothers in it. This is because the 20–30 bar is twice as wide. Both bars have the same density of mothers per cigarette (i.e., 3.2% per cigarette).

Histograms can also be used to answer distributional questions such as: what proportion of the babies weigh under 100 ounces or what percentage of the babies weigh more than 138 ounces. From the histogram in Figure 1.5, we sum the areas of the bars to the left of 100 and find that 14% of the babies weigh under 100

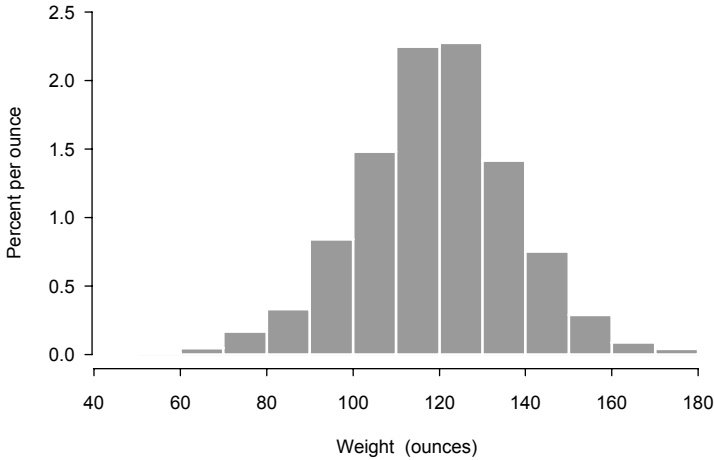


FIGURE 1.5. Histogram of infant birth weight for 1236 babies in the CHDS subset.

ounces. However, to answer the second question, we note that 138 does not fall at an interval endpoint of the histogram, so we need to approximate how many babies weigh between 138 and 140 ounces. To do this, split up the interval that runs from 130 to 140 into 10 one-ounce subintervals. The bar contains 14.2% of the babies, so we estimate that each one-ounce subinterval contains roughly 1.4% of the babies and 2.8% of the babies weigh 138–140 ounces. Because 12.5% of the babies weigh over 140 ounces, our estimate is that 15.3% of the babies weigh more than 138 ounces. In fact, 15.1% of the babies weighed more than this amount. The approximation was quite good.

Numerical Summaries

A measure of location is a statistic that represents the center of the data distribution. One such measure is the mean, which is the average of the data. The mean can be interpreted as the balance point of the histogram. That is, if the histogram were made up of bars sitting on a weightless balance beam, the mean would be the point at which the histogram would balance on the beam.

For a list of numbers x_1, \dots, x_n , the mean \bar{x} is computed as follows:

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i.$$

A measure of location is typically accompanied by a measure of dispersion that gives an idea as to how far an individual value may vary from the center of the

data. One such measure is the standard deviation (SD). The standard deviation is the root mean square (r.m.s.) of the deviations of the numbers on the list from the list average. It is computed as

$$SD(x) = \sqrt{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2}.$$

An alternative measure of location is the median. The median is the point that divides the data (or list of numbers) in half such that at least half of the data are smaller than the median and at least half are larger. To find the median, the data must be put in order from smallest to largest.

The measure of dispersion that typically accompanies the median is the interquartile range (IQR). It is the difference between the upper and lower quartiles of the distribution. Roughly, the lower quartile is that number such that at least 25% of the data fall at or below it and at least 75% fall at or above it. Similarly, the upper quartile is the number such that at least 75% of the data fall at or below it and at least 25% fall at or above it. When more than one value meets this criterion, then typically the average of these values is used. For example, with a list of 10 numbers, the median is often reported as the average of the 5th and 6th largest numbers, and the lower quartile is reported as the 3rd smallest number.

For infant birth weight, the mean is 120 ounces and the SD is 18 ounces. Also, the median is 120 ounces and the IQR is 22 ounces. The mean and median are very close due to the symmetry of the distribution. For heavily skewed distributions, they can be very far apart. The mean is easily affected by outliers or an asymmetrically long tail.

Five-Number Summary

The five-number summary provides a measure of location and spread plus some additional information. The five numbers are: the median, the upper and lower quartiles, and the extremes (the smallest and largest values). The five-number summary is presented in a box, such as in Table 1.4, which is a five-number summary for the weights of 1200 mothers in the CHDS.

From this five-number summary, it can be seen that the distribution of mother's weight seems to be asymmetric. That is, it appears to be either skewed to the right or to have some large outliers. We see this because the lower quartile is closer to the median than the upper quartile and because the largest observation is very far

TABLE 1.4. Five-number summary for the weights (in pounds) of 1200 mothers in the CHDS subset.

Median	125	
Quartiles	115	139
Extremes	87	250

from the upper quartile. Half of the mothers weigh between 115 and 139 pounds, but at least one weighs as much as 250 pounds.

Box-and-Whisker Plot

A box-and-whisker plot is another type of graphical representation of data. It contains more information than a five-number summary but not as much information as a histogram. It shows location, dispersion and outliers, and it may indicate skewness and tail size. However, from a box-and-whisker plot it is not possible to ascertain whether there are gaps or multiple modes in a distribution.

In a box-and-whisker plot, the bottom of the box coincides with the lower quartile and the top with the upper quartile; the median is marked by a line through the box; the whiskers run from the quartiles out to the smallest (largest) number that falls within $1.5 \times \text{IQR}$ of the lower (upper) quartile; and smaller or larger numbers are marked with a special symbol such as a * or —.

Figure 1.6 contains a box-and-whisker plot of mother's weight. The right skewness of the distribution is much more apparent here than in the five-number summary. There are many variants on the box-and-whisker plot, including one that simply draws whiskers from the sides of the box to the extremes of the data.

The Normal Curve

The standard normal curve (Figure 1.7), known as the bell curve, sometimes provides a useful method for summarizing data.

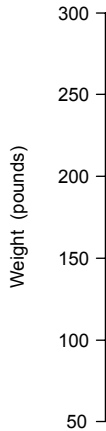


FIGURE 1.6. Box-and-whisker plot of mother's weight for 1200 mothers in the CHDS subset.

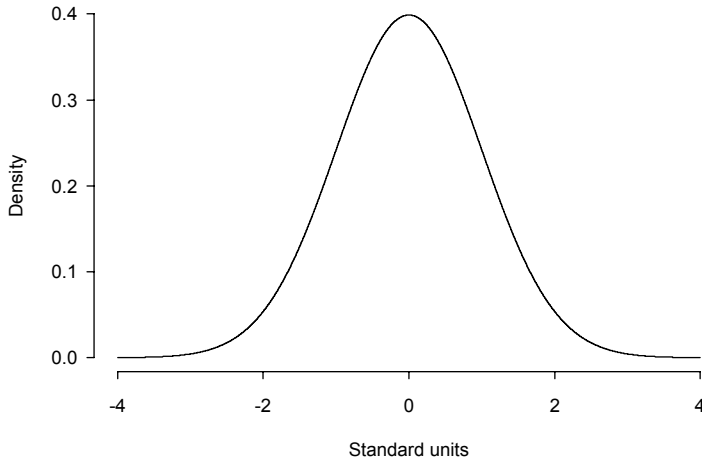


FIGURE 1.7. The standard normal curve.

The normal curve is unimodal and symmetric around 0. It also follows the 68-95-99.7 rule. The rule states that 68% of the area under the curve is within 1 unit of its center, 95% is within 2 units of the center, and 99.7% is within 3 units of its center. These areas and others are determined from the following analytic expression for the curve:

$$\frac{1}{\sqrt{2\pi}} e^{-x^2/2}.$$

Traditionally, $\Phi(z)$ represents the area under the normal curve to the left of z , namely,

$$\Phi(z) = \int_{-\infty}^z \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx.$$

A table of these areas can be found in Appendix C. Also, most statistical software provides these numbers.

Many distributions for data are approximately normal, and the 68-95-99.7 rule can be used as an informal check of normality. If the histogram looks normal, then this rule should roughly hold when the data are properly standardized. Note that to standardize the data, subtract the mean from each number and then divide by the standard deviation; that is, compute

$$\frac{x_i - \bar{x}}{\text{SD}(x)}.$$

Notice that a value of +1 for the standard normal corresponds to an x -value that is 1 SD above \bar{x} . We saw in Figure 1.2 that standardizing the birth weights of babies led to a more informative comparison of mortality rates for smokers and nonsmokers.

For birth weight, we find that 69% of the babies have weights within 1 standard deviation of the average, 96% are within 2 SDs, and 99.4% are within 3 SDs. It looks pretty good. When the normal distribution fits well and we have summarized the data by its mean and SD, the normal distribution can be quite handy for answering such questions as what percentage of the babies weigh more than 138 ounces. The area under the normal curve can be used to approximate the area of the histogram. When standardized, 138 is 1 standard unit above average. The area under a normal curve to the right of 1 is 16%. This is close to the actual figure of 15%.

Checks for normality that are more formal than the 68-95-99.7 rule are based on the coefficients of skewness and kurtosis. In standard units, the coefficient of skewness is the average of the third power of the standardized data, and the coefficient of kurtosis averages the 4th power of the standardized list. That is,

$$\text{skewness} = \frac{1}{n} \sum_{i=1}^n \left(\frac{x_i - \bar{x}}{\text{SD}(x)} \right)^3 \quad \text{kurtosis} = \frac{1}{n} \sum_{i=1}^n \left(\frac{x_i - \bar{x}}{\text{SD}(x)} \right)^4 .$$

For a symmetric distribution, the skewness coefficient is 0. The kurtosis is a measure of how pronounced is the peak of the distribution. For the normal, the kurtosis should be 3. Departures from these values (0 for skewness and 3 for kurtosis) indicate departures from normality.

To decide whether a given departure is big or not, simulation studies can be used. A simulation study generates pseudo-random numbers from a known distribution, so we can check the similarity between the simulated observations and the actual data. This may show us that a particular distribution would be unlikely to give us the data we see. For example, the kurtosis of birth weight for the 484 babies born to smokers in the CHDS subset is 2.9. To see if 2.9 is a typical kurtosis value for a sample of 484 observations from a normal distribution, we could repeat the following a large number of times: generate 484 pseudo-random observations from a normal distribution and calculate the sample kurtosis. Figure 1.8 is a histogram of 1000 sample values of kurtosis computed for 1000 samples of size 484 from the standard normal curve. From this figure, we see that 2.9 is a very typical kurtosis value for a sample of 484 from a standard normal.

Quantile Plots

For a distribution such as the standard normal, the q th quantile is z_q , where

$$\Phi(z_q) = q, \quad 0 < q < 1.$$

The median, lower, and upper quartiles are examples of quantiles. They are, respectively, the 0.50, 0.25, and 0.75 quantiles.

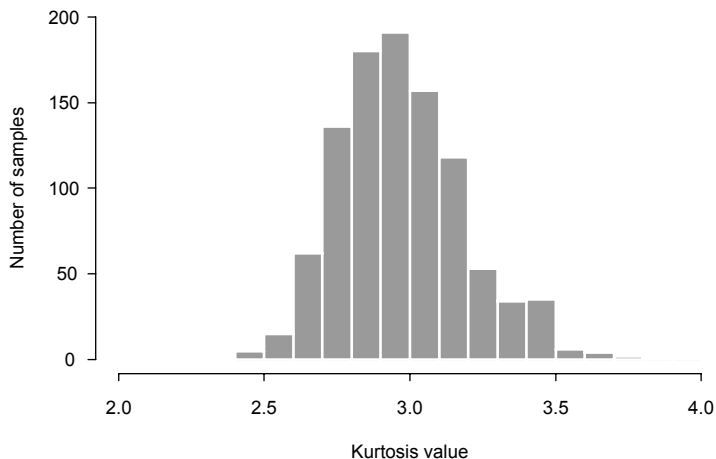


FIGURE 1.8. Histogram of kurtosis values for 1000 samples of size 484 from the standard normal.

For data x_1, \dots, x_n , the sample quantiles are found by ordering the data from smallest to largest. We denote this ordering by $x_{(1)}, \dots, x_{(n)}$. Then $x_{(k)}$ is considered the $k/(n+1)$ th sample quantile. We divide by $n+1$ rather than n to keep q less than 1.

The normal-quantile plot, also known as the normal-probability plot, provides a graphical means of comparing the data distribution to the normal. It graphs the pairs $(z_{k/(n+1)}, x_{(k)})$. If the plotted points fall roughly on a line, then it indicates that the data have an approximate normal distribution. See the Exercises for a more formal treatment of quantiles. Figure 1.9 is a normal-quantile plot of the weights of mothers in the CHDS. The upward curve in the plot identifies a long right tail, in comparison to the normal, for the weight distribution.

Departures from normality are indicated by systematic departures from a straight line. Examples of different types of departures are provided in Figure 1.10. Generally speaking, if the histogram of the data does not decrease as quickly in the right tail as the normal, this is indicated by an upward curve on the right side of the normal-quantile plot. Similarly, a long left tail is indicated by a downward curve to the left (bottom right picture in Figure 1.10). On the other hand, if the tails decrease more quickly than the normal, then the curve will be as in the bottom left plot in Figure 1.10. Granularity in the recording of the data appears as stripes in the plot (top left plot in Figure 1.10). Bimodality is shown in the top right plot of Figure 1.10.

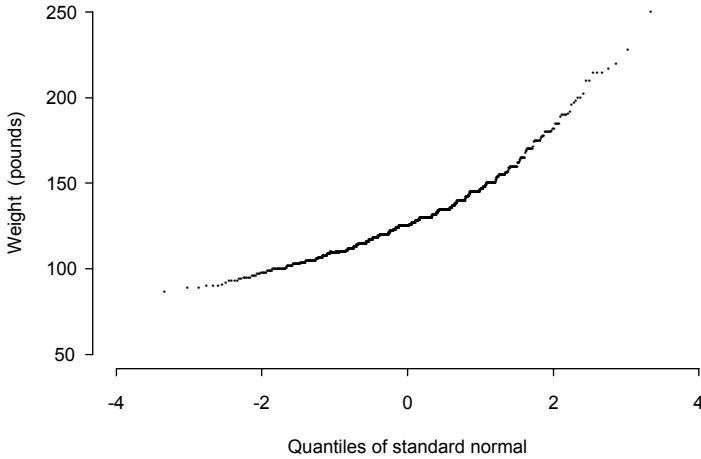


FIGURE 1.9. Normal quantile plot of mother's weight for 1200 mothers in the CHDS subset.

Quantile plots can be made for any distribution. For example, a uniform-quantile plot for mother's weight appears in Figure 1.11, where the sample quantiles of mother's weight are plotted against the quantiles of the uniform distribution. It is evident from the plot that both the left and right tails of the weight distribution are long in comparison to the uniform.

To compare two data distributions — such as the weights of smokers and non-smokers — plots known as quantile-quantile plots can be made. They compare two sets of data to each other by pairing their respective sample quantiles. Again, a departure from a straight line indicates a difference in the shapes of the two distributions. When the two distributions are identical, the plot should be linear with slope 1 and intercept 0 (roughly speaking, of course). If the two distributions are the same shape but have different means or standard deviations, then the plot should also be roughly linear. However, the intercept and slope will not be 0 and 1, respectively. A nonzero intercept indicates a shift in the distributions, and a nonunit slope indicates a scale change. Figure 1.12 contains a quantile-quantile plot of mother's weight for smokers and nonsmokers compared with a line of slope 1 and intercept 0. Over most of the range there appears to be linearity in the plot, though lying just below the line: smokers tend to weigh slightly less than nonsmokers. Notice that the right tail of the distribution of weights is longer for the nonsmokers, indicating that the heaviest nonsmokers weigh quite a bit more than the heaviest smokers.

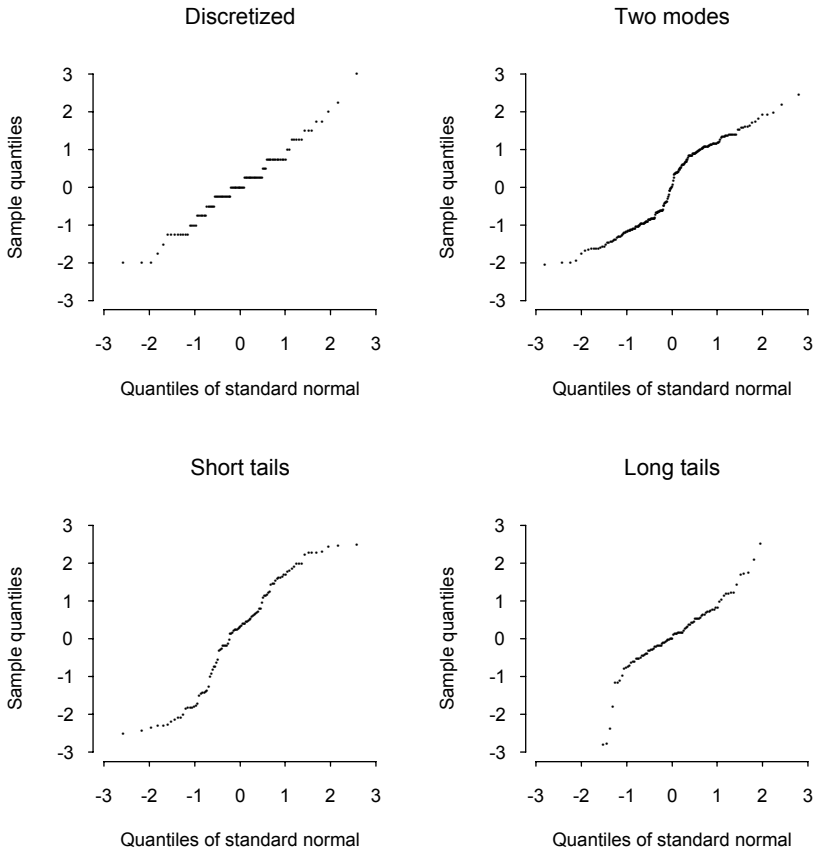


FIGURE 1.10. Examples of normal quantile plots.

Cross-tabulations

Distribution tables for subgroups of the data are called cross-tabulations. They allow for comparisons of distributions across more homogeneous subgroups. For example, the last row of Table 1.5 contains the distribution of body length for a sample of 663 babies from the CHDS. The rows of the table show the body-length distribution for smokers and nonsmokers separately. Notice that the babies of the smokers seem to be shorter than the babies of nonsmokers. It looks as though the distribution for the smokers is shifted to the left.

Bar Charts and Segmented Bar Charts

A bar chart is often used as a graphical representation of a cross-tabulation. It depicts the count (or percent) for each category of a second variable within each

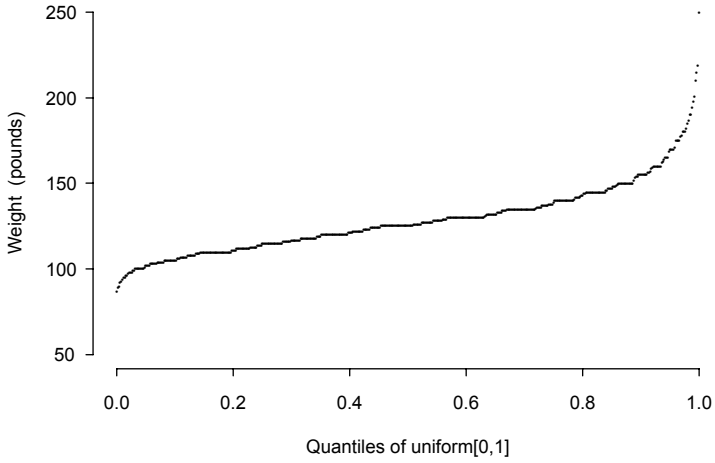


FIGURE 1.11. Uniform-quantile plot of mother's weight for 1200 mothers in the CHDS subset.

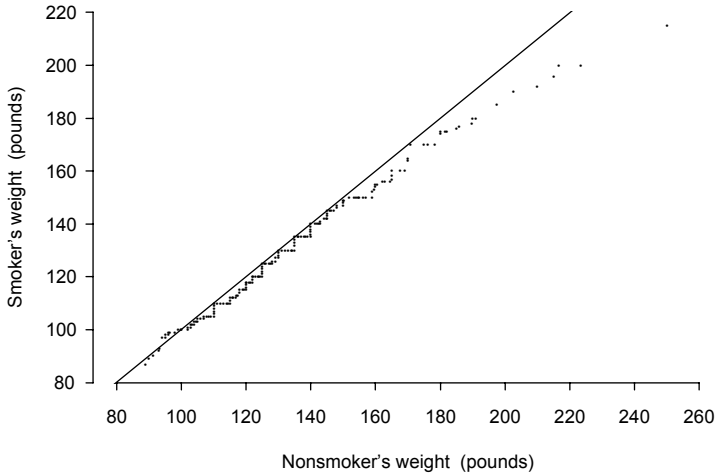


FIGURE 1.12. Quantile-quantile plot of mother's weight for smokers (484) and nonsmokers (752) in the CHDS subset; superimposed is a line of slope 1 and intercept 0.

TABLE 1.5. Cross-tabulation of infant body length (in inches) for smokers and nonsmokers for a sample of 663 babies from the CHDS.

		Body length (inches)					Total
		≤18	19	20	21	≥22	
Nonsmokers	Count	18	70	187	175	50	500
	Percent	4	14	37	35	10	100
Smokers	Count	5	42	56	47	13	163
	Percent	3	26	34	29	8	100
Total count		23	112	243	222	63	663

TABLE 1.6. Population characteristics and prevalence of maternal smoking among 305,730 births to white Missouri residents, 1979–1983 (Malloy et al. [MKLS88]).

		Percent of mothers	Percent smokers in each group
All		100	30
Marital status	Married	90	27
	Single	10	55
Educational level (years)	Under 12	21	55
	12	46	29
	Over 12	33	15
Maternal age (years)	Under 18	5	43
	18–19	9	44
	20–24	35	34
	25–29	32	23
	30–34	15	21
	Over 34	4	26

category of a first variable. A segmented bar chart stacks the bars of the second variable, so that their total height is the total count for the category of the first variable (or 100 percent). Table 1.6 contains comparisons of smokers and nonsmokers according to marital status, education level, and age. The segmented bar chart in the left plot of Figure 1.13 shows the percentage of unmarried and married mothers who are smokers and nonsmokers. This information can also be summarized where one bar represents the smokers, one bar represents the nonsmokers, and the shaded region in a bar denotes the proportion of unmarried mothers in the group (6% for nonsmokers and 19% for smokers). Alternatively, a bar chart of these data might show the shaded and unshaded bars adjacent to each other rather than stacked. (These alternative figures are not depicted).

Table 10.3 in Chapter 10 compares qualitative characteristics of the families in the CHDS study according to whether the mother smokes or not. One of these characteristics, whether the mother uses contraceptives or not, is pictured in the segmented bar chart in the right plot of Figure 1.13.

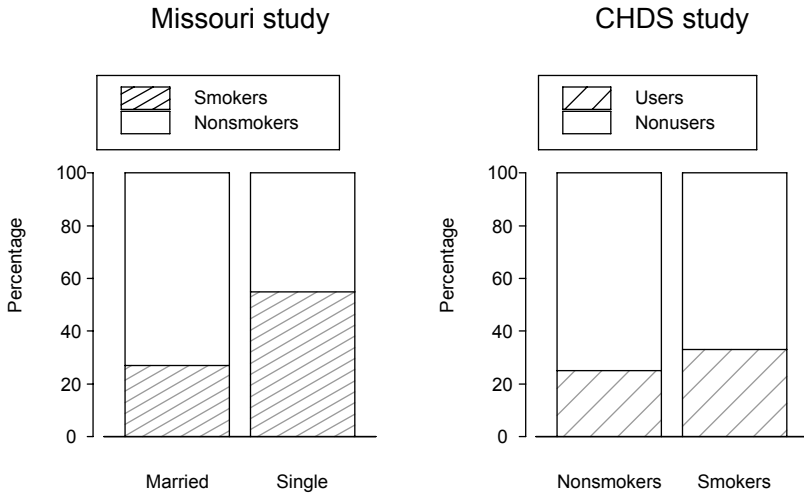


FIGURE 1.13. Bar charts of smoking prevalence by marital status (left) for mothers in the Missouri study (Malloy et al. [MKLS88]) and contraceptive use by smoking prevalence (right) for mothers in the CHDS study (Yerushalmy [Yer71]).

Exercises

1. Use Table 1.3 to find the approximate quartiles of the distribution of the number of cigarettes smoked per day for the mothers in the CHDS who smoked during their pregnancy.
2. Combine the last four categories in Table 1.3 of the distribution of the number of cigarettes smoked by the smoking mothers in the CHDS. Make a new histogram using the collapsed table. How has the shape changed from the histogram in Figure 1.4? Explain.
3. Consider the histogram of father's age for the fathers in the CHDS (Figure 1.14). The bar over the interval from 35 to 40 years is missing. Find its height.
4. Consider the normal quantile plots of father's height and weight for fathers in the CHDS (Figure 1.15). Describe the shapes of the distributions.
5. Following are the quantiles at 0.05, 0.10, ..., 0.95 for the gestational ages of the babies in the CHDS. Plot these quantiles against those of the uniform distribution on (0, 1). Describe the shape of the distribution of gestational age in comparison to the uniform.
252, 262, 267, 270, 272, 274, 276, 277, 278, 280, 281, 283, 284, 286, 288, 290, 292, 296, 302.

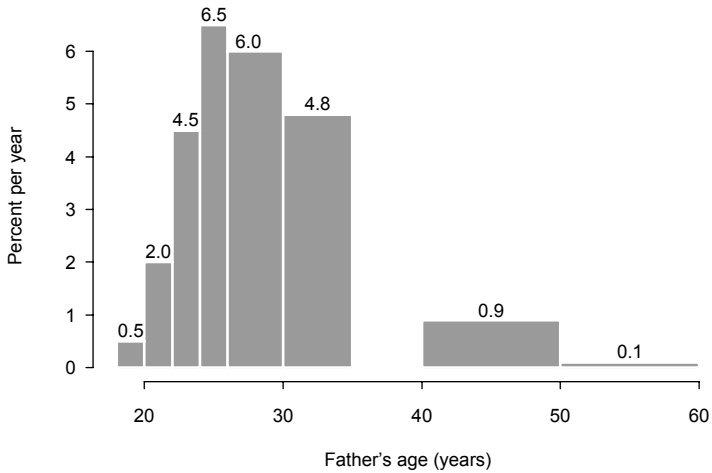


FIGURE 1.14. Histogram of father's age for fathers in the CHDS, indicating height of the bars. The bar over the interval from 35 to 40 years is missing.

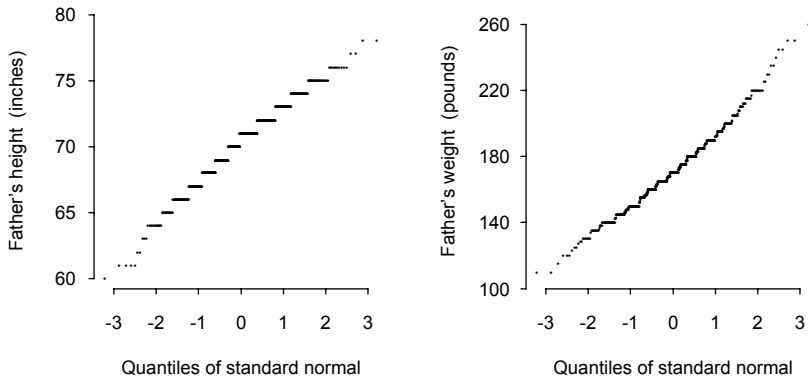


FIGURE 1.15. Normal quantile plots of father's height (left) and weight (right) for fathers in the CHDS.

- Use the normal approximation to estimate the proportion of mothers in the CHDS between 62 and 64 inches tall to the nearest half inch (i.e., between 61.5 and 64.5 inches). The average height is 64 inches and the SD is 2.5 inches.

7. In the Missouri study, the average birth weight for babies born to smokers is 3180 grams and the SD 500 grams, and for nonsmokers the average is 3500 grams and the SD 500 grams. Consider a baby who is born to a smoker. If the baby's weight is 2 SDs below average weighs, then the baby weighs _____ grams. Suppose another baby weighs this same number of grams, but is born to a nonsmoker. This baby has a weight that falls _____ SDs below the average of its group. According to the normal approximation, approximately what percentage of babies born to nonsmokers are below this weight?
8. Suppose there are 100 observations from a standard normal distribution. What proportion of them would you expect to find outside the whiskers of a box-and-whisker plot?
9. Make a table for marital status that gives the percentage of smokers and nonsmokers in each marital category for the mothers in the Missouri study (Table 1.6).
10. Make a segmented bar graph showing the percentage at each education level for both smokers and nonsmokers for the mothers in the Missouri study (Table 1.6).
11. Make a bar graph of age and smoking status for the mothers in the Missouri study (Table 1.6). For each age group, the bar should denote the percentage of mothers in that group who smoke. How are age and smoking status related? Is age a potential confounding factor in the relationship between a mother's smoking status and her baby's birth weight?
12. In the Missouri study, the average birth weight for babies born to smokers is 3180 grams and the SD is 500 grams. What is the average and SD in ounces? There are 0.035 ounces in 1 gram.
13. Consider a list of numbers x_1, \dots, x_n . Shift and rescale each x_i as follows:

$$y_i = a + bx_i.$$

Find the new average and SD of the list y_1, \dots, y_n in terms of the average and SD of the original list x_1, \dots, x_n .

14. Consider the data in Exercise 13. Express the median and IQR of y_1, \dots, y_n in terms of the median and IQR of x_1, \dots, x_n . For simplicity, assume $y_1 < y_2 < \dots < y_n$ and assume n is odd.
15. For a list of numbers x_1, \dots, x_n with $x_1 < x_2 < \dots < x_n$, show that by replacing x_n with another number, the average and SD of the list can be made arbitrarily large. Is the same true for the median and IQR? Explain.
16. Suppose there are n observations from a normal distribution. How could you use the IQR of the list to estimate σ ?
17. Suppose the quantiles y_q of a $\mathcal{N}(\mu, \sigma^2)$ distribution are plotted against the quantiles z_q of a $\mathcal{N}(0, 1)$ distribution. Show that the slope and intercept of the line of points are σ and μ , respectively.
18. Suppose X_1, \dots, X_n form a sample from the standard normal. Show each of the following:

- a. $\Phi(X_1), \dots, \Phi(X_n)$ is equivalent to a sample from a uniform distribution on $(0, 1)$. That is, show that for X a random variable with a standard normal distribution,

$$\mathbb{P}(\Phi(X) \leq q) = q.$$

- b. Let U_1, \dots, U_n be a sample from a uniform distribution on $(0, 1)$. Explain why

$$\mathbb{E}(U_{(k)}) = \frac{k}{n+1}.$$

where $U_{(1)} \leq \dots \leq U_{(n)}$ are the ordered sample.

- c. Use (a) and (b) to explain why $X_{(k)} \approx z_{k/n+1}$.
19. Prove that \bar{x} is the constant that minimizes the following squared error with respect to c :

$$\sum_{i=1}^n (x_i - c)^2.$$

20. Prove that the median \tilde{x} of x_1, \dots, x_n is the constant that minimizes the following absolute error with respect to c :

$$\sum_{i=1}^n |x_i - c|.$$

You may assume that there are an odd number of distinct observations. *Hint:* Show that if $c < c_o$, then

$$\sum_{i=1}^n |x_i - c_o| = \sum_{i=1}^n |x_i - c| + (c - c_o)(r - s) + 2 \sum_{x \in (c, c_o)} (c - x),$$

where $r = \text{number of } x_i \geq c_o$, and $s = n - r$.

Notes

Yerushalmy's original analysis of the CHDS data ([Yer64], [Yer71]) and Hodges et al. ([HKC75]) provide the general framework for the analysis found in this lab and its second part in Chapter 10.

The data for the lab are publicly available from the School of Public Health at the University of California at Berkeley. Brenda Eskanazi and David Lein of the School of Public Health provided valuable assistance in the extraction of the data used in this lab.

The information on fetal development is adapted from Samuels and Samuels ([SS86]).

References

- [Gre41] N.M. Gregg. Congenital cataract following German measles in the mother. *Trans. Ophthalmol. Soc. Aust.*, **3**:35–46, 1941.
- [HKC75] J.L. Hodges, D. Krech, and R.S. Crutchfield. *Instructor's Handbook to Accompany StatLab*. McGraw–Hill Book Company, New York, 1975.
- [Lon76] L. Longo. Carbon monoxide: Effects on oxygenation of the fetus in utero. *Science*, **194**: 523–525, 1976.
- [MKLS88] M. Malloy, J. Kleinman, G. Land, and W. Schram. The association of maternal smoking with age and cause of infant death. *Am. J. Epidemiol.*, **128**:46–55, 1988.
- [MT77] M.B. Meyer and J.A. Tonascia. Maternal smoking, pregnancy complications, and perinatal mortality. *Am. J. Obstet. Gynecol.*, **128**: 494–502, 1977.
- [MT90] I. Merkatz and J. Thompson. *New Perspectives on Prenatal Care*. Elsevier, New York, 1990.
- [SS86] M. Samuels and N. Samuels. *The Well Pregnancy Book*. Summit Books, New York, 1986.
- [Wil93] A.J. Wilcox. Birthweight and perinatal mortality: The effect of maternal smoking. *Am. J. Epidemiol.*, **137**:1098–1104, 1993.
- [WR86] A.J. Wilcox and I.T. Russell. Birthweight and perinatal mortality, III: Towards a new method of analysis. *Int. J. Epidemiol.*, **15**:188–196, 1986.
- [Yer64] J. Yerushalmy. Mother's cigarette smoking and survival of infant. *Am. J. Obstet. Gynecol.*, **88**:505–518, 1964.
- [Yer71] J. Yerushalmy. The relationship of parents' cigarette smoking to outcome of pregnancy—implications as to the problem of inferring causation from observed associations. *Am. J. Epidemiol.*, **93**:443–456, 1971.