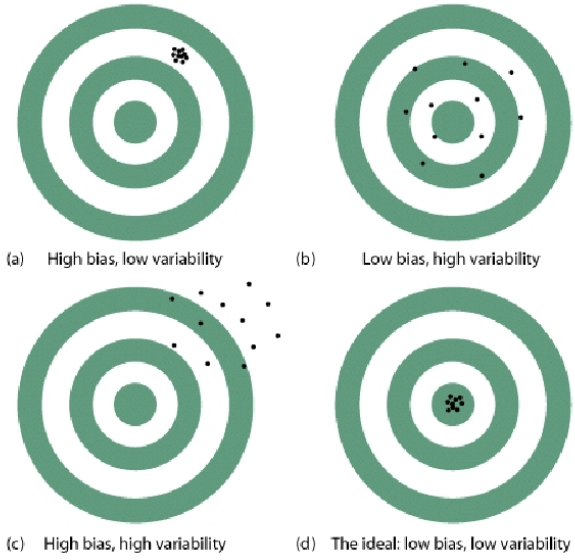
CHAPTER 8

FUNDAMENTAL SAMPLING DISTRIBUTIONS AND DATA DESCRIPTIONS

8.1 Random Sampling

The basic idea of the statistical inference is that we are allowed to draw inferences or conclusions about a population based on the statistics computed from the sample data so that we could infer something about the parameters and obtain more information about the population. Thus we must make sure that **the samples must be good representatives of the population** and pay attention on the sampling bias and variability to ensure the validity of statistical inference.



Bias

Any sampling procedure that produces inferences that consistently overestimate or consistently underestimate some characteristic of the population is said to be *biased*.

To eliminate any possibility of bias in the sampling procedure, it is desirable to choose a **random sample** in the sense that the observations are made independently and at random.

Random Sample

Let $X_1, X_2, \ldots, X_n$ be n independent random variables, each having the same probability distribution $f(x)$. Define $X_1, X_2, \ldots, X_n$ to be a random sample of size n from the population $f(x)$ and write its joint probability distribution as

$$f(x_1, x_2, \ldots, x_n) = f(x_1) f(x_2) \cdots f(x_n).$$

8.2 Some Important Statistics

It is important to measure the center and the variability of the population. For the purpose of the inference, we study the following measures regarding to the center and the variability.

8.2.1 Location Measures of a Sample

The most commonly used statistics for measuring the center of a set of data, arranged in order of magnitude, are the **sample mean**, **sample median**, and **sample mode**. Let $X_1, X_2, \ldots, X_n$ represent n random variables.

Sample Mean

To calculate the average, or *mean*, add all values, then divide by the number of individuals.

$$\overline{X} = \frac{X_1 + X_2 + \cdots + X_n}{n} = \frac{1}{n} \sum_{i=1}^{n} X_i$$

where $\overline{X}$ is the special symbol of the sample mean and $\overline{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$ denotes its value, or the realization of X.

NOTE. The mean is the *balance* point. It is the "center of mass".

EXAMPLE 8.1. The weights of a group of students (in lbs) are given below:

135 105 118 163 172 183 122 150 121 162

Find the mean. If another student joins in the group and his weight is 250 lbs, what would be the new mean?

Sample Median

The number such that half of the observations are smaller and half are larger, i.e., the *midpoint* of a distribution.

$$\tilde{x} = \begin{cases} x_{(n+1)/2} & \text{if } n \text{ is odd} \\ \frac{1}{2}\left(x_{n/2} + x_{n/2+1}\right) & \text{if } n \text{ is even} \end{cases}$$

EXAMPLE 8.2. The weights of a group of students (in lbs) are given below:

135 105 118 163 172 183 122 150 121 162

Find the median. If another student joins in the group and his weight is 250 lbs, what would be the new median?

Sample Mode

The *mode* of a data set is the value that occurs *most frequently*.

The cases are unimodal, bimodal, multimodal and no mode. The mode is/are the value(s) whose frequencies are the largest (the peaks).

EXAMPLE 8.3. The weights of three group of students (in lbs) are given below:

(a) $135, 105, 118, 163, 172, 183, 122, 150$

(b) $135, 105, 118, 163, 172, 183, 122, 135$

(c) $135, 135, 118, 118, 122, 118, 122, 135$

Find the mode for each group.

8.2.2 Variability Measures of a Sample

The most commonly used statistics for measuring the center of a set of data, arranged in order of magnitude, are the **sample variance**, **sample standard deviation**, and **sample range**. Let $X_1, X_2, \ldots, X_n$ represent n random variables.

Sample Variance

The sample *variance* "S^2" is used to describe the variation around the mean. We use

$$\begin{aligned} s^2 &= \frac{1}{n-1}\sum (x_i - \bar{x})^2 \\ &= \frac{1}{n-1}\left[\sum x^2 - \frac{(\sum x)^2}{n}\right] \\ &= \frac{n\sum(x^2) - (\sum x)^2}{n(n-1)} \end{aligned}$$

to denote the realization or the computed values of S^2.

Sample Standard Deviation

The sample *standard deviation* is the squared root of the sample variance.

$$S = \sqrt{S^2}$$

and

$$s = \sqrt{\frac{1}{n-1}\sum (x_i - \bar{x})^2}$$

NOTE. Properties of Standard Deviation

- s measures spread about the mean and should be used only when the mean is the measure of center.
- $s = 0$ only when all observations have the same value and there is no spread. Otherwise, $s > 0$.
- s gets larger, as the observations become more spread out about their mean.
- s has the same units of measurement as the original observations.

NOTE. The standard deviation of a population is defined by $\sigma = \sqrt{\frac{\sum(x-\mu)^2}{N}}$, where N is the population size and μ is population mean. *Be careful with the denominator inside square-root is N, instead of $N-1$.*

EXAMPLE 8.4. Calculate the sample variance and the sample standard deviation of the following set of data:

$$0 \quad 1 \quad -2 \quad -3 \quad 9$$

Sample Range

The sample *range* R of a data set is defined as

$$R = X_{\max} - X_{\min}$$

EXAMPLE 8.5. Refer to Example 8.4. Find the sample range.

8.3 Sampling Distributions

Sampling Distribution

In general, the *sampling distribution* of a given statistic is the distribution of the values taken by the statistic in all possible samples of the same size form the same population.

In other words, if we repeatedly collect samples of the same sample size from the population, compute the statistics (mean, standard deviation, proportion), and then draw a histogram of those statistics, the distribution of that histogram tends to have is called the **sample distribution** of that statistics (mean, standard deviation, proportion).

NOTE. The statistical applets are good tools to study the sampling distribution. Check out the Rice University Applets at http://onlinestatbook.com/stat_sim/sampling_dist/index.html.

8.4 Sampling Distribution of Means and the Central Limit Theorem

8.4.1 Sampling Distribution of Sample Means from a Normal Population

Mean and Standard Deviation of a Sample Mean

Theorem. Let $\overline{X}$ be the sample mean of a random sample of size n drawn from a population having mean μ and standard deviation σ, then the mean of $\overline{X}$ is

$$\mu_{\overline{X}} = \mu$$

and the standard deviation of $\overline{X}$ is

$$\sigma_{\overline{X}} = \frac{\sigma}{\sqrt{n}}$$

EXAMPLE 8.6. Prove the above theorem.

The mean and standard deviation of $\bar{x}$

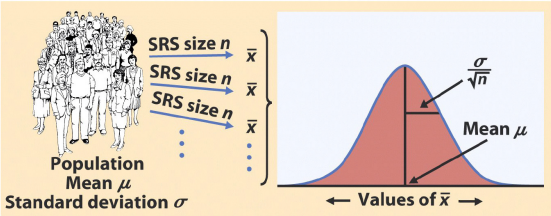


Figure 5-12
Introduction to the Practice of Statistics, Fifth Edition

NOTE.

- The sample mean $\overline{X}$ is an *unbiased* estimator of the population mean μ and is less variable than a single observation.
- The variation of $\overline{X}$ is much smaller than that of the population. The standard deviation of $\overline{X}$ decreases as the sample size n increases.
- The above results do NOT require any assumptions on the shape of the population. However, a random sample is a must.

EXAMPLE 8.7. The mean and standard deviation of the strength of a packaging material are 55 kg and 6 kg, respectively. A quality manager takes a random sample of specimens of this material and tests their strength. If the manager wants to reduce the standard deviation of $\overline{X}$ to 1.5 kg, how many specimens should be tested?

EXAMPLE 8.8. A soft-drink machine is regulated so that the amount of drink dispensed averages 240 milliliters with a standard deviation of 15 milliliters. Periodically, the machine is checked by taking a sample of 40 drinks and computing the average content. If the mean of the 40 drinks is a value within the interval $\mu_{\overline{X}} \pm 2\sigma_{\overline{X}}$, the machine is thought to be operating satisfactorily; otherwise, adjustments are made. The company official found the mean of 40 drinks to be $\bar{x} = 236$ milliliters and concluded that the machine needed no adjustment. Was this a reasonable decision?

Sampling Distribution of Sample Means from a Normal Population

Theorem. Let $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ be the sample mean of a random sample of size n drawn from a **normal** population having mean μ and standard deviation σ, then $\overline{X}$ follows an *exact* **normal** distribution with mean μ and standard deviation $\sigma/\sqrt{n}$. That is,

$$X_i \sim \mathrm{N}(\mu,\, \sigma) \quad \Longrightarrow \quad \overline{X} \sim \mathrm{N}\left(\mu,\, \sigma/\sqrt{n}\right).$$

EXAMPLE 8.9. Prove the above theorem.

NOTE.

- One of the essential assumptions is **a random sample**.
- The distribution of $\overline{X}$ has the *EXACTLY normal* distribution if the random sample is from a **normal population**.

EXAMPLE 8.10. The contents of bottles of beer vary according to a normal distribution with mean $\mu = 341$ ml and standard deviation $\sigma = 3$ ml.

(a) What is the probability that the content of a randomly selected bottle is less than 339 ml?

(b) What is the probability that the average content of the bottles in a 12-pack of beer is less than 339 ml?

EXAMPLE 8.11. A patient is classified as having gestational diabetes if the glucose level is above 140 milligrams per deciliter (mg/dl) one hour after a sugary drink is ingested. Sheila's measured glucose level one hour after ingesting the sugary drink varies according to the normal distribution with $\mu = 125$ mg/dl and $\sigma = 10$ mg/dl.

(a) If a single glucose measurement is made, what is the probability that Sheila is diagnosed as having gestational diabetes?

(b) If measurements are made on three separate days and the mean result is compared with the criterion 140 mg/dl, what is the probability that Sheila is diagnosed as having gestational diabetes?

(c) What is the level L such that there is probability only 5% that the mean glucose level of three test results fall above L for Sheila's glucose level distribution.

8.4.2 The Central Limit Theorem (CLT)

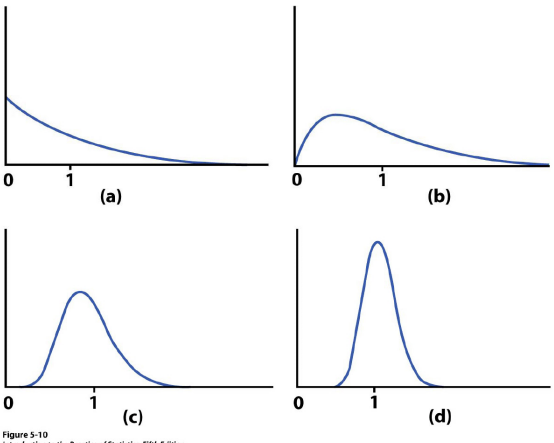


Figure 5-10
Introduction to the Practice of Statistics, Fifth Edition
© 2005 W. H. Freeman and Company

Theorem (Central Limit Theorem)**.** If $\overline{X}$ is the mean of a random sample of size n taken from a population with mean μ and finite variance σ^2, then

$$Z = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \to \mathrm{n}(z; 0, 1)$$

as $n \to \infty$.

In other words, if a random sample of size n is selected from **any** population with mean μ and standard deviation σ, then

$$\overline{X} \text{ is approximately } \mathrm{N}\left(\mu,\ \sigma/\sqrt{n}\right),$$

when n is *sufficiently large*.

NOTE. The Central Limit Theorem is important because, for reasonably large sample size, it allows us to make an approximate probability statement concerning the sample mean, without knowledge of the shape of the population distribution.

- Again, one of the essential assumptions is *a random sample*.
- The distribution of $\overline{X}$ has the *approximately normal* distribution if the random sample is from a population *other than* normal.
- How large a sample size? Usually, it would safe to apply the CLT if $n \geq 30$. It also depends on the population distribution, however. More observations are required if the population distribution is far from normal.

EXAMPLE 8.12. The time a family physician spends seeing a patient follows some right-skewed distribution with a mean of 15 minutes and a standard deviation of 11.6 minutes.

(a) Can you calculate the probability that the doctor spends less than 12 minutes with the next patient she sees? If so, do it. If not, explain why.

(b) What is the probability that the doctor spends an average time between 13 and 18 minutes with her 30 patients of the day?

(c) One day, 35 patients have an appointment to see the doctor. What is the probability that she will have to work overtime, beyond her 8-hour shift?

8.4.3 Sampling Distribution of the Difference between Two Means

Suppose that we have two populations, the first with mean μ_1 and standard deviation σ_1, and the second with mean μ_2 and standard deviation σ_2. We take a random sample of size n_1 from the first population and measure some variable X_1, and take an **independent** random sample of size n_2 from the second population and measure the value of the some variable X_2.

By the Central Limit Theorem, we know that, if n_1 and n_2 are *sufficiently large*,

$$\overline{X}_1 \ \dot{\sim}\ \mathrm{N}(\mu_1,\ \sigma_1/\sqrt{n_1}),$$

and

$$\overline{X}_2 \ \dot{\sim}\ \mathrm{N}(\mu_2,\ \sigma_2/\sqrt{n_2}).$$

It can also be shown that

$$\overline{X}_1 - \overline{X}_2 \ \dot{\sim}\ \mathrm{N}\left(\mu_1 - \mu_2,\ \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}\right).$$

Theorem. If independent samples of size n_1 and n_2 are drawn at random from two populations, discrete or continuous, with means μ_1 and μ_2 and variances σ_1^2 and σ_2^2, respectively, then the sampling distribution of the differences of means, $\overline{X}_1 - \overline{X}_2$, is approximately normally distributed with mean and variance given by

$$\mu_{\overline{X}_1-\overline{X}_2} = \mu_1 - \mu_2 \quad \text{and} \quad \sigma^2_{\overline{X}_1-\overline{X}_2} = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}.$$

So,

$$Z = \frac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}} \;\dot{\sim}\; \mathrm{N}(0,1)$$

NOTE. If both samples are from the normal populations, the sampling distribution of $\overline{X}_1 - \overline{X}_2$ will be **exactly normal**, instead of approximate normal.

EXAMPLE 8.13. We take a random sample of five 10-year-old boys and four 10-year-old girls and measure their heights. Suppose that we know that heights X_1 of 10-year old boys follow a normal distribution with mean 55.7 inches and standard deviation 2.9 inches, and that heights X_2 of 10-year old girls follow a normal distribution with mean 54.1 inches and standard deviation 2.6 inches. What is the probability that the mean height of the girls in the sample is smaller than the mean height for the boys in the sample?

EXAMPLE 8.14. A research on bulimia among college women studies the connection between childhood sexual abuse and a measure of family cohesion (the higher the score, the greater the cohesion). Assume that sexually abused students have an average family cohesion scale of 2.8 and a standard deviation of 2.1, while non-abused students have the average scale of 4.8 and a standard deviation of 3.2. What is the probability that a random sample of 49 non-abused students will have an average family cohesion scale that is at least 0.5 scores higher than the average scale of a randoms sample of 36 sexually abused students? What can you conclude?

8.5 Sampling Distribution of S^2

Distribution of $(n-1)S^2/\sigma^2$

If S^2 is the variance of a random sample of size n taken from a normal population having the variance σ^2, then the statistic

$$\chi^2 = \frac{(n-1)S^2}{\sigma^2} = \sum_{i=1}^{n} \frac{(X_i - \overline{X})^2}{\sigma^2}$$

has a chi-squared distribution with $\nu = n-1$ degrees of freedom.

NOTE (Degrees of Freedom). There are n degrees of freedom, or independent pieces of information, in the random sample from the normal distribution. When the data (the values in the sample) are used to compute the mean (i.e., when μ is replaced by $\bar{x}$), a degree of freedom is lost in the estimation of μ. Hence, there are the remaining $(n-1)$ degrees of freedom in the information used to estimate σ^2.

Let $\chi^2_\alpha(\nu)$ be the χ^2 value above which we find an area of α under the curve of the chi-squared distribution with ν degrees of freedom. That is,

$$\mathrm{P}\left(\chi^2(\nu) > \chi^2_\alpha(\nu)\right) = \alpha.$$

We use table A.5. to find these critical values of the chi-squared distribution with ν degrees of freedom.

EXAMPLE 8.15. Find the critical values

(a) $\chi^2_{0.95}(4)$

(b) $\chi^2_{0.75}(22)$

EXAMPLE 8.16. Find k such that $\mathrm{P}\left(\chi^2(12) < k\right) = 0.80$.

EXAMPLE 8.17. Use Table A.5. to give the best estimate to each of the following probabilities.

(a) $\mathrm{P}\left(\chi^2(5) \geq 3\right)$

(b) $\mathrm{P}\left(\chi^2(8) > 3.33\right)$

(c) $\mathrm{P}\left(\chi^2(10) \leq 6.66\right)$

(d) $\mathrm{P}\left(\chi^2(25) > 99.9\right)$

8.6 t-Distribution

We have learned that $Z = \frac{\overline{X}-\mu}{\sigma/\sqrt{n}}$ (exactly or approximately) follows the standard normal distribution, where the data are from a *random sample* of size n from the population with mean μ and standard deviation σ. And, it is very likely that both μ and σ are unknown parameters. In practice, it suffices that the distribution is symmetric and single-peaked unless the sample is very small.

Since most of the simple work in statistical inference focus on the unknown population mean μ, we will need deal with the unknown σ especially when n is not large. It is quite intuitive and natural to estimate the unknown population standard deviation σ using the sample standard deviation S.

We have another statistic $T = \dfrac{\overline{X}-\mu}{S/\sqrt{n}}$ as an analog sample version of $Z = \dfrac{\overline{X}-\mu}{\sigma/\sqrt{n}}$.

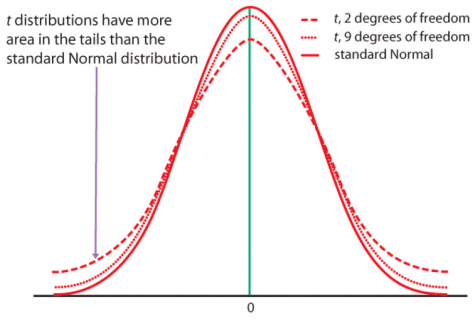
Student t distribution

Let $X_1, X_2, \ldots, X_n$ be independent random variables that are all normal with mean μ and standard deviation σ. Let $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ and $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2$. Then the random variable

$$T = \frac{\overline{X} - \mu}{S/\sqrt{n}}$$

has a t-distribution with $\nu = n-1$ degrees of freedom.

NOTE. When n is very large, s is a very good estimate of σ, and the corresponding t distributions are very close to the normal distribution. The t distributions become wider for smaller sample sizes, reflecting the lack of precision in estimating σ from s.

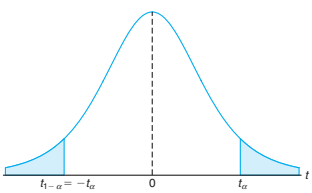


Important Properties of the Student t distribution

- The t distribution is different for different sample sizes, or different degrees of freedom.
- The t distribution has the same general symmetric bell shape as the standard normal distribution, but it reflects the greater variability (with wider distributions) that is expected with small samples.
- The t distribution has a mean of $t = 0$.
- The standard deviation of the t distribution varies with the sample size, but it is greater than 1.
- As the sample size n gets larger, the t distribution gets closer to the standard normal distribution.

Let $t_\alpha(\nu)$ be the t value above which we find an area of α under the curve of the t distribution with ν degrees of freedom. That is,

$$\mathrm{P}\left(T(\nu) > t_\alpha(\nu)\right) = \alpha.$$



Because the symmetrically property, $t_{1-\alpha} = -t_\alpha$.

We use table A.4. to find these critical values of the t distribution with ν degrees of freedom.

NOTE. The t table, as well as the χ^2 table, gives us the **UPPER** tail probabilities, while the z table gives the *lower* tail probabilities.

EXAMPLE 8.18. Find the critical values.

(a) $t_{0.005}(5)$, $t_{0.05}(5)$, $t_{0.5}(5)$, $t_{0.85}(5)$, $t_{0.975}(5)$

(b) $t_{0.10}(10)$, $t_{0.20}(20)$, $t_{0.30}(30)$, $t_{0.40}(40)$, $t_{0.60}(60)$

(c) $t_{0.90}(10)$, $t_{0.95}(15)$, $t_{0.99}(19)$

EXAMPLE 8.19. Let $T(\nu)$ denote the Student t-distribution with ν degrees of freedom. Find k such that

(a) $\mathrm{P}(T(8) > k) = 0.02$.

(b) $\mathrm{P}(T(18) < k) = 0.80$.

(c) $\mathrm{P}(T(28) \geq k) = 0.99$.

EXAMPLE 8.20. Use Table A.4. to give the best estimate to each of the following probabilities.

(a) $\mathrm{P}(T(5) \geq 1.11)$

(b) $\mathrm{P}(T(8) < 2.22)$

(c) $\mathrm{P}(T(10) \geq 3.33)$

(d) $\mathrm{P}(T(15) > 4.44)$

NOTE. Clearly,

$$T = \frac{\frac{\overline{X}-\mu}{\sigma/\sqrt{n}}}{\sqrt{\frac{\left[(n-1)S^2/\sigma^2\right]}{n-1}}} = \frac{Z}{\sqrt{\frac{\chi^2(n-1)}{n-1}}}$$

More generally,

Theorem. Let Z be a standard normal random variable and V a chi-squared random variable with ν degrees of freedom. If Z and V are independent, then the distribution of the random variable T, where

$$T = \frac{Z}{\sqrt{V/\nu}}$$

is given by the density function

$$h(t) = \frac{\Gamma[(\nu+1)/2]}{\Gamma(\nu/2)\sqrt{\pi\nu}}\left(1 + \frac{t^2}{\nu}\right)^{-(\nu+1)/2}, \quad -\infty < t < \infty.$$

This is known as the t-distribution with ν degrees of freedom.