确定课程论文选题：

分组：每组3-4人合作完成，提交纸质报告并做口头报告。每个小组在期中需要选定一个小课题，要求调研至少一种课上没有讲到的数据分析或图像处理方法，讨论其在科研（不限于天文）中的应用状况，用真实数据重复他人结果或自己做出新的分析，并在课堂上讲解其背景、原理、分析结果。

Survey sampling

Opening the gate to mathematical statistics

Sample surveys

are used to obtain information about a large population by examining only a small fraction of that population.

- Governments survey human populations, conduct health surveys and census surveys.
- In agriculture, to estimate such quantities as the total acreage of wheat in a state by surveying a sample of farms.
- Sampling studies of rail and highway traffic. In one such study, records of shipments of household goods by motor carriers were sampled to evaluate the accuracy of preshipment estimates of charges, claims for damages, and other variables.
- In the practice of quality control, the output of a manufacturing process may be sampled in order to examine the items for defects.
- During audits (审计) of the financial records of large companies or institutions (e.g. USTC!), sampling techniques may be used when examination of the entire set of records is impractical.

Sample surveys

Probabilistic in nature -- *each member has a specified probability of being included in the sample, and the actual composition of the sample is random.*

Sampling schemes in contrast: particular members are included in the sample because they are thought to be typical in some way – may be effective in some situations, but there is no way mathematically to guarantee its unbiasedness or to estimate the magnitude of any error committed.

Sample surveys

Probabilistic in nature -- *each member has a specified probability of being included in the sample, and the actual composition of the sample is random.*

Sampling schemes in contrast: particular members are included in the sample because they are thought to be typical in some way – may be effective in some situations, but there is no way mathematically to guarantee its unbiasedness or to estimate the magnitude of any error committed.

- The selection of sample units at random is a guard against investigator biases, even unconscious ones. （无偏）

Sample surveys

Probabilistic in nature -- *each member has a specified probability of being included in the sample, and the actual composition of the sample is random.*

Sampling schemes in contrast: particular members are included in the sample because they are thought to be typical in some way – may be effective in some situations, but there is no way mathematically to guarantee its unbiasedness or to estimate the magnitude of any error committed.

- The selection of sample units at random is a guard against investigator biases, even unconscious ones. （无偏）
- A small sample costs far less and is much faster to survey than a complete enumeration. （快捷低成本）

Sample surveys

Probabilistic in nature -- *each member has a specified probability of being included in the sample, and the actual composition of the sample is random.*

Sampling schemes in contrast: particular members are included in the sample because they are thought to be typical in some way – may be effective in some situations, but there is no way mathematically to guarantee its unbiasedness or to estimate the magnitude of any error committed.

- The selection of sample units at random is a guard against investigator biases, even unconscious ones. （无偏）
- A small sample costs far less and is much faster to survey than a complete enumeration. （快捷低成本）
- Results from a small sample ***may actually be more accurate!!!*** than those from a complete enumeration – **Why?**

Sample surveys

Probabilistic in nature -- *each member has a specified probability of being included in the sample, and the actual composition of the sample is random.*

Sampling schemes in contrast: particular members are included in the sample because they are thought to be typical in some way – may be effective in some situations, but there is no way mathematically to guarantee its unbiasedness or to estimate the magnitude of any error committed.

- The selection of sample units at random is a guard against investigator biases, even unconscious ones. （无偏）
- A small sample costs far less and is much faster to survey than a complete enumeration. （快捷低成本）
- Results from a small sample ***may actually be more accurate*** than those from a complete enumeration -- data quality more easily monitored and controlled, a complete enumeration may require a much larger (and thus more poorly trained) staff. （更精准）

Sample surveys

Probabilistic in nature -- *each member has a specified probability of being included in the sample, and the actual composition of the sample is random.*

Sampling schemes in contrast: particular members are included in the sample because they are thought to be typical in some way – may be effective in some situations, but there is no way mathematically to guarantee its unbiasedness or to estimate the magnitude of any error committed.

- The selection of sample units at random is a guard against investigator biases, even unconscious ones. （无偏）
- A small sample costs far less and is much faster to survey than a complete enumeration. （快捷低成本）
- Results from a small sample ***may actually be more accurate*** than those from a complete enumeration -- data quality more easily monitored and controlled, a complete enumeration may require a much larger (and thus more poorly trained) staff. （更精准）
- Random sampling techniques make possible the calculation of an estimate of the error due to sampling. （可算误差）

Sample surveys

Probabilistic in nature -- *each member has a specified probability of being included in the sample, and the actual composition of the sample is random.*

Sampling schemes in contrast: particular members are included in the sample because they are thought to be typical in some way – may be effective in some situations, but there is no way mathematically to guarantee its unbiasedness or to estimate the magnitude of any error committed.

- The selection of sample units at random is a guard against investigator biases, even unconscious ones. （无偏）
- A small sample costs far less and is much faster to survey than a complete enumeration. （快捷低成本）
- Results from a small sample ***may actually be more accurate*** than those from a complete enumeration -- data quality more easily monitored and controlled, a complete enumeration may require a much larger (and thus more poorly trained) staff. （更精准）
- Random sampling techniques make possible the calculation of an estimate of the error due to sampling. （可算误差）
- In designing a sample, it is frequently possible to determine the sample size necessary to obtain a prescribed error level. （按需设计）

Population parameters

Numerical characteristics, or parameters, of the population that we will estimate from a sample. Population size = N, numerical values x_1, x_2,…, x_N, are age, weight, …, or dichotomous (0 or 1).

Example: Herkson (1976). The population consists of N = 393 short-stay hospitals. Let x_i denote the number of patients discharged from the ith hospital during January 1968. See histogram: # of hospitals that discharged 0-200, 201-400,…, 2801-3000 patients were plotted.

Population mean:

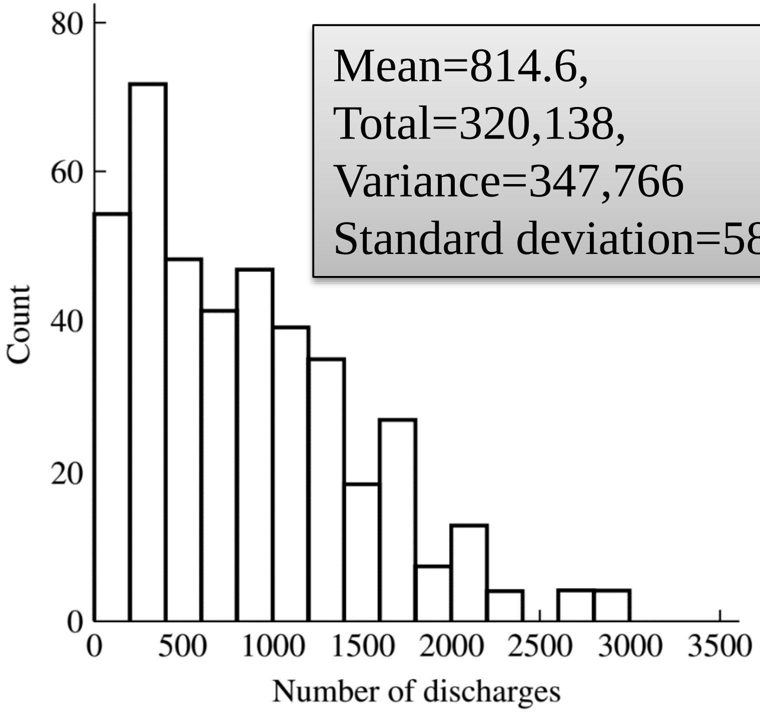
$$\mu = \frac{1}{N}\sum_{i=1}^{N} x_i$$

Population total:

$$\tau = \sum_{i=1}^{N} x_i = N\mu$$

Population variance:

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2$$


Mean=814.6,
Total=320,138,
Variance=347,766
Standard deviation=589.7

Population parameters

Dichotomous (0 or 1) case:

Population mean = proportion *p*, of individuals having the particular characteristic.
Population total = total # of members possessing the characteristic of interest.

$$\sigma^2 = \frac{1}{N}\left(\sum_{i=1}^{N} x_i^2 - 2\mu \sum_{i=1}^{N} x_i + N\mu^2\right)$$

$$= \frac{1}{N}\left(\sum_{i=1}^{N} x_i^2 - 2N\mu^2 + N\mu^2\right)$$

$$= \frac{1}{N}\sum_{i=1}^{N} x_i^2 - \mu^2$$

Population variance:

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N} x_i^2 - \mu^2$$

$$= p - p^2$$

$$= p(1-p)$$

… and population standard deviation is also defined.

Simple random sampling (简单随机抽样)

The most elementary form is s.r.s.

Each particular sample of size *n* has the same probability of occurrence -- each of the $\binom{N}{n}$ possible samples of size *n* taken without replacement has the same probability.

- We assume no replacement so that each member appears in the sample at most once (e.g. balls in an urn, no replacement).

Sample composition is random

$\Rightarrow$ sample mean is random

$\Rightarrow$ accuracy analysis where sample mean approximates population mean is probabilistic in nature.

$\Rightarrow$ Let's look at sample mean…

Sample Mean: Its expectation and variance

Population size=*N*, sample size=*n*, values of sample members X_1, X_2,…, X_N.

Sample mean as an estimate of population mean:

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

Estimate of population total:

$$T = N\overline{X}$$

Sample mean (and others) has so-called “**sampling distribution**”.

Sample Mean: Its expectation and variance

Population size=N, sample size=n, values of sample members X_1, X_2,…, X_N.
Sample mean as an estimate of population mean:

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

Estimate of population total:

$$T = N\overline{X}$$

Sample mean (and others) has so-called “**sampling distribution**”.

Example. 393 hospitals again. We want to find the sampling distribution of the mean of a sample of size 16. In principle, we could form all $\binom{393}{16}$ samples, compute the mean per each, find the sampling distribution (~10^{28}!!).

Sample Mean: Its expectation and variance

Population size=N, sample size=n, values of sample members X_1, X_2,…, X_N. Sample mean as an estimate of population mean:

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

Estimate of population total:

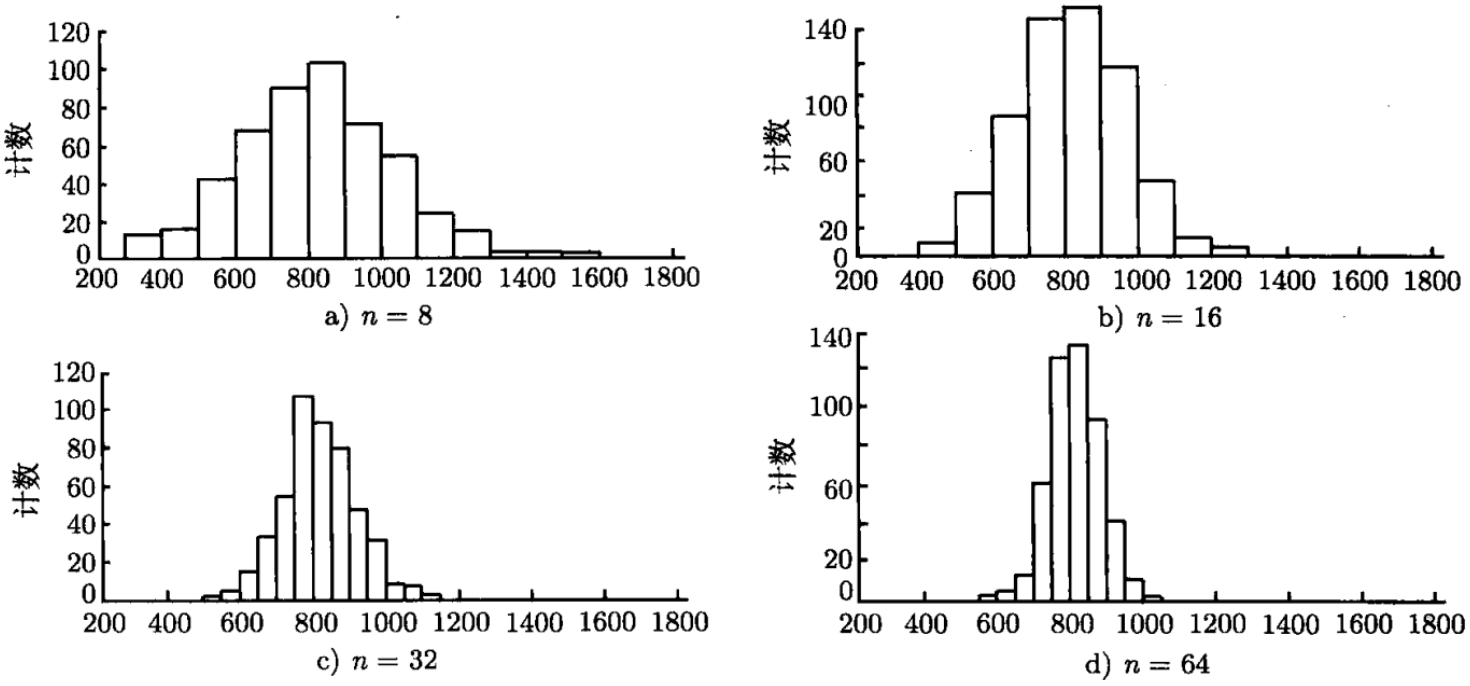
$$T = N\overline{X}$$

Sample mean (and others) has so-called "**sampling distribution**".

Example. 393 hospitals again. We want to find the sampling distribution of the mean of a sample of size 16. In principle, we could form all $\binom{393}{16}$ samples, compute the mean per each, find the sampling distribution (~10^{28}!!).

Employ a technique known as **simulation**: estimate the sampling distribution of the mean of a sample of size n by drawing many samples of size n, computing the mean of each sample, and then forming a histogram of the collection of sample means.

Sample Mean: Its expectation and variance



a) $n = 8$

b) $n = 16$

c) $n = 32$

d) $n = 64$

Sample sizes of 8, 16, 32, 64, with 500 replications for each sample size.

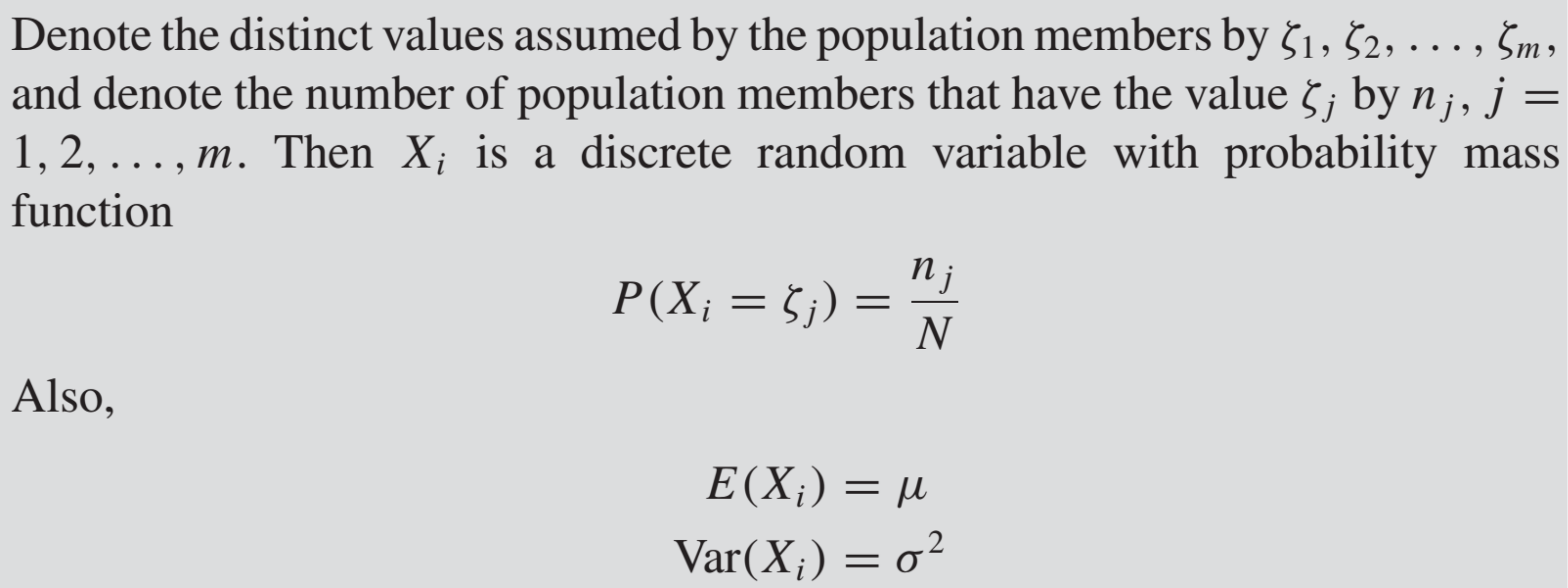
1. All the histograms are centered about the population mean, 814.6.
2. As the sample size increases, the histograms become less spread out.
3. Although the shape of the histogram of population values is not symmetric about the mean, the histograms are more nearly so.

Sample Mean: Its expectation and variance

If all members are distinct, $P(X_i=x_j)=1/N$.

If not distinct (e.g. dichotomous: 0 or 1), k members are same, then $P(X_i=\zeta)=k/N$.

With or without replacement, we have a lemma:



Denote the distinct values assumed by the population members by $\zeta_1, \zeta_2, \ldots, \zeta_m$, and denote the number of population members that have the value ζ_j by n_j, $j = 1, 2, \ldots, m$. Then X_i is a discrete random variable with probability mass function

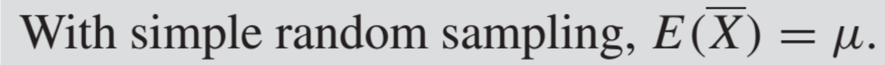
$$P(X_i = \zeta_j) = \frac{n_j}{N}$$

Also,

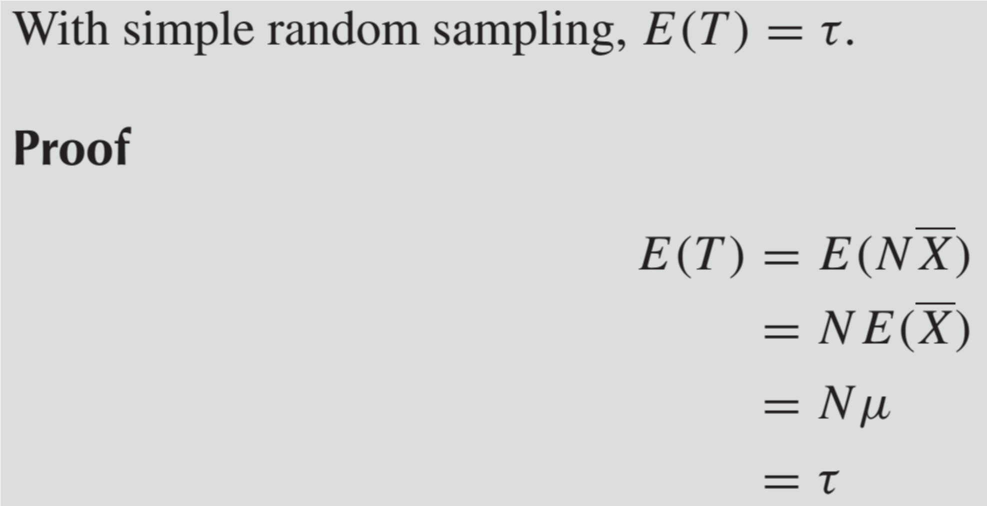
$$E(X_i) = \mu$$
$$\mathrm{Var}(X_i) = \sigma^2$$

Sample Mean: Its expectation and variance

The sampling distribution is centered at μ, a theorem:



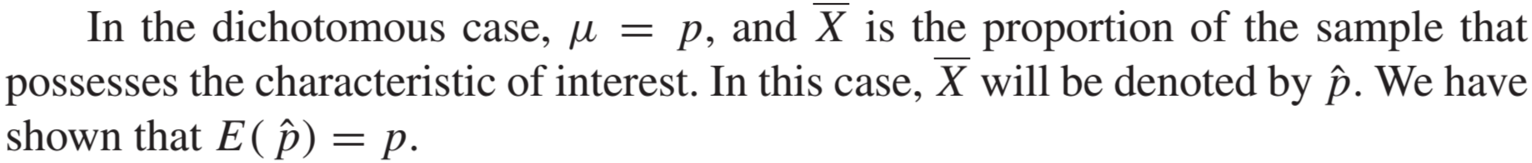
With simple random sampling, $E(\overline{X}) = \mu$.



With simple random sampling, $E(T) = \tau$.

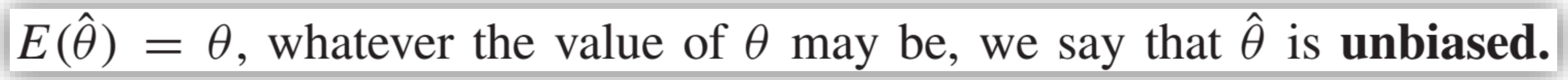
Proof

$$\begin{aligned} E(T) &= E(N\overline{X}) \\ &= NE(\overline{X}) \\ &= N\mu \\ &= \tau \end{aligned}$$

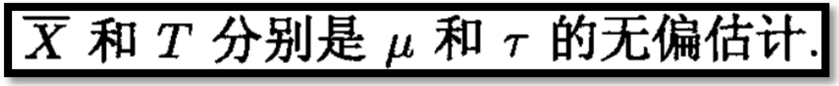


In the dichotomous case, $\mu = p$, and $\overline{X}$ is the proportion of the sample that possesses the characteristic of interest. In this case, $\overline{X}$ will be denoted by $\hat{p}$. We have shown that $E(\hat{p}) = p$.

If we estimate a population parameter θ, by a function of the sample, then if



$E(\hat{\theta}) = \theta$, whatever the value of θ may be, we say that $\hat{\theta}$ is **unbiased.**



$\overline{X}$ 和 T 分别是 μ 和 τ 的无偏估计.

Sample Mean: Its variance

We next find $\mathrm{Var}(\overline{X})$. Since $\overline{X} = n^{-1}\sum_{i=1}^{n} X_i$,

$$\mathrm{Var}(\overline{X}) = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\mathrm{Cov}(X_i,\ X_j)$$

If the sampling were done with replacement (可重复抽样), then X_i are independent, for $i \neq j$ we would have $\mathrm{Cov}(X_i, X_j) = 0$, whereas $\mathrm{Cov}(X_i, X_i) = \mathrm{Var}(X_i) = \sigma^2$.

$$\mathrm{Var}\,\overline{X} = \frac{1}{n^2}\sum_{i=1}^{n}\mathrm{Var}\,(X_i) = \frac{\sigma^2}{n} \qquad \sigma_{\overline{X}} = \frac{\sigma}{\sqrt{n}}$$

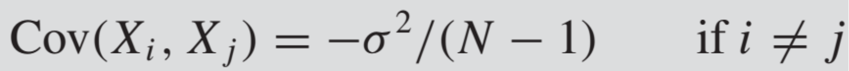
standard error

Sampling without replacement **introduces dependence**, though if n<<N then this results holds to a good approximation.

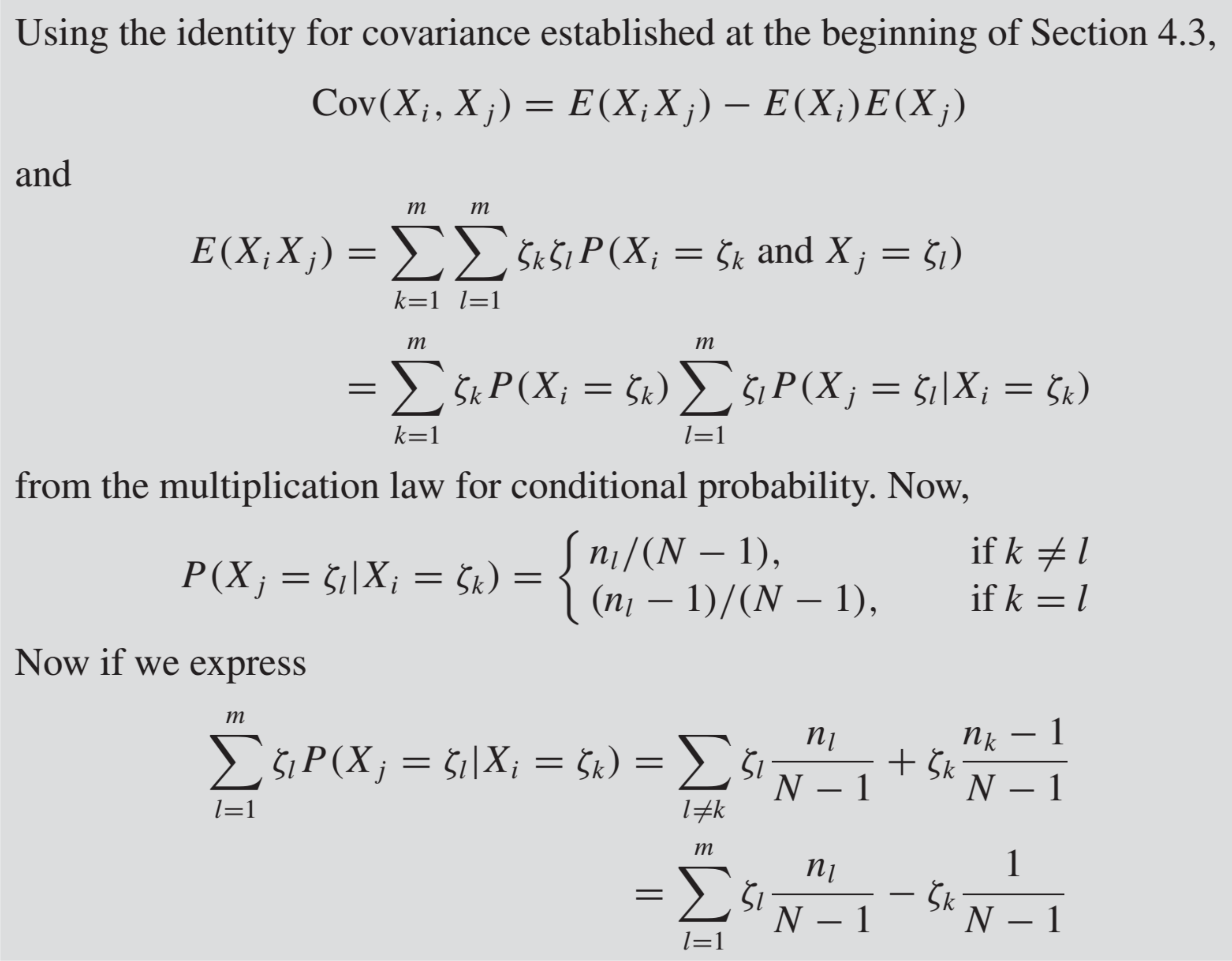
For simple random sampling without replacement,

$$\mathrm{Cov}(X_i, X_j) = -\sigma^2/(N-1) \qquad \text{if } i \neq j$$

Proof.



$$\mathrm{Cov}(X_i, X_j) = -\sigma^2/(N-1) \qquad \text{if } i \neq j$$



Using the identity for covariance established at the beginning of Section 4.3,

$$\mathrm{Cov}(X_i, X_j) = E(X_i X_j) - E(X_i)E(X_j)$$

and

$$\begin{aligned} E(X_i X_j) &= \sum_{k=1}^{m} \sum_{l=1}^{m} \zeta_k \zeta_l P(X_i = \zeta_k \text{ and } X_j = \zeta_l) \\ &= \sum_{k=1}^{m} \zeta_k P(X_i = \zeta_k) \sum_{l=1}^{m} \zeta_l P(X_j = \zeta_l | X_i = \zeta_k) \end{aligned}$$

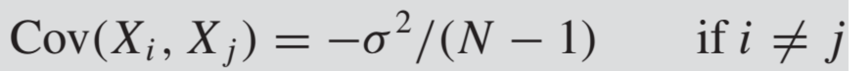
from the multiplication law for conditional probability. Now,

$$P(X_j = \zeta_l | X_i = \zeta_k) = \begin{cases} n_l/(N-1), & \text{if } k \neq l \\ (n_l - 1)/(N-1), & \text{if } k = l \end{cases}$$

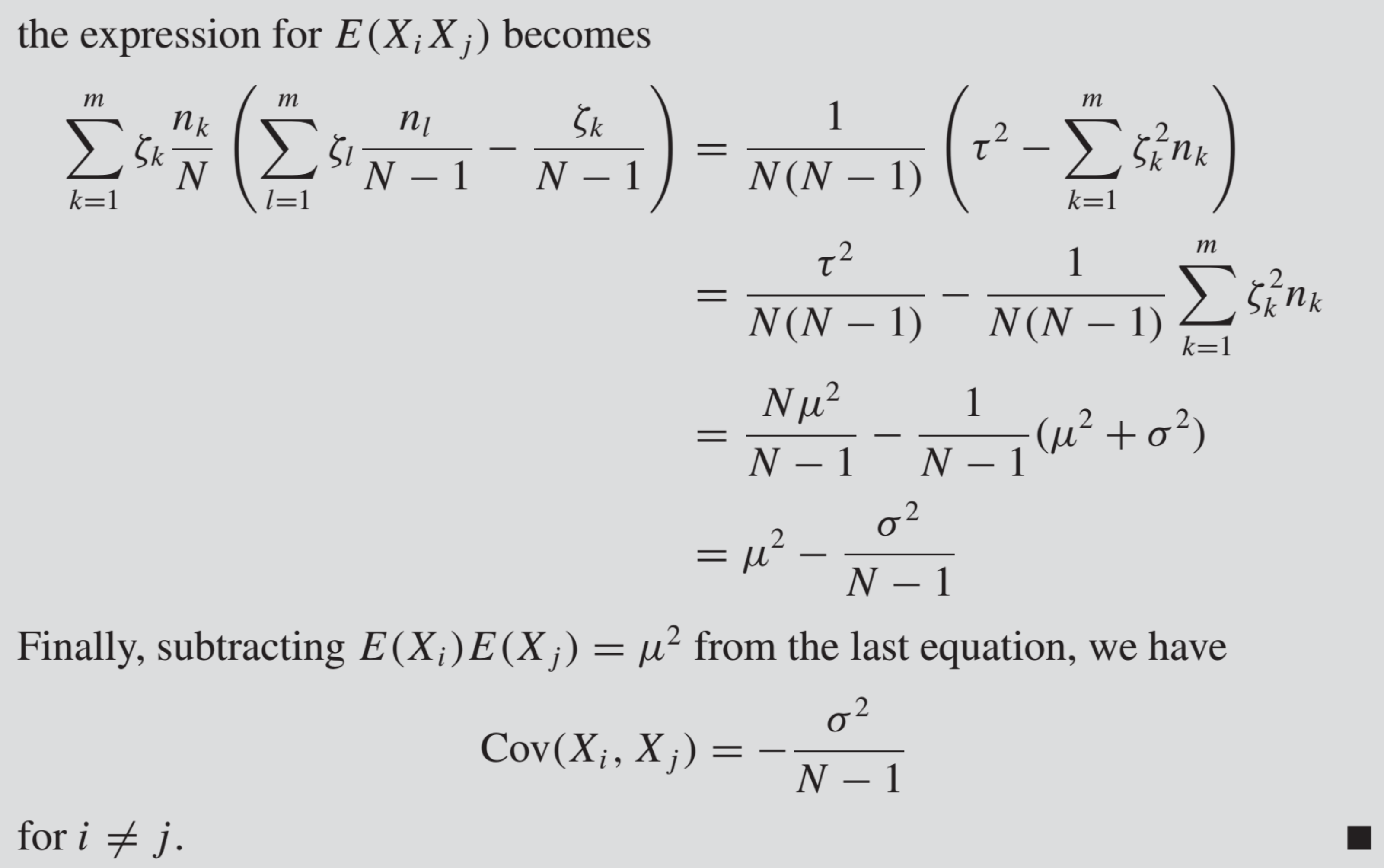
Now if we express

$$\begin{aligned} \sum_{l=1}^{m} \zeta_l P(X_j = \zeta_l | X_i = \zeta_k) &= \sum_{l \neq k} \zeta_l \frac{n_l}{N-1} + \zeta_k \frac{n_k - 1}{N-1} \\ &= \sum_{l=1}^{m} \zeta_l \frac{n_l}{N-1} - \zeta_k \frac{1}{N-1} \end{aligned}$$

Proof.



$$\mathrm{Cov}(X_i, X_j) = -\sigma^2/(N-1) \qquad \text{if } i \neq j$$



the expression for $E(X_i X_j)$ becomes

$$\begin{aligned}
\sum_{k=1}^{m} \zeta_k \frac{n_k}{N} \left(\sum_{l=1}^{m} \zeta_l \frac{n_l}{N-1} - \frac{\zeta_k}{N-1} \right) &= \frac{1}{N(N-1)} \left(\tau^2 - \sum_{k=1}^{m} \zeta_k^2 n_k \right) \\
&= \frac{\tau^2}{N(N-1)} - \frac{1}{N(N-1)} \sum_{k=1}^{m} \zeta_k^2 n_k \\
&= \frac{N\mu^2}{N-1} - \frac{1}{N-1}(\mu^2 + \sigma^2) \\
&= \mu^2 - \frac{\sigma^2}{N-1}
\end{aligned}$$

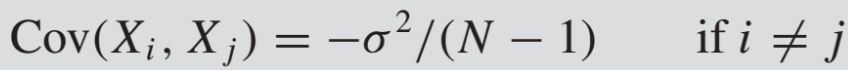
Finally, subtracting $E(X_i)E(X_j) = \mu^2$ from the last equation, we have

$$\mathrm{Cov}(X_i, X_j) = -\frac{\sigma^2}{N-1}$$

for $i \neq j$. ■

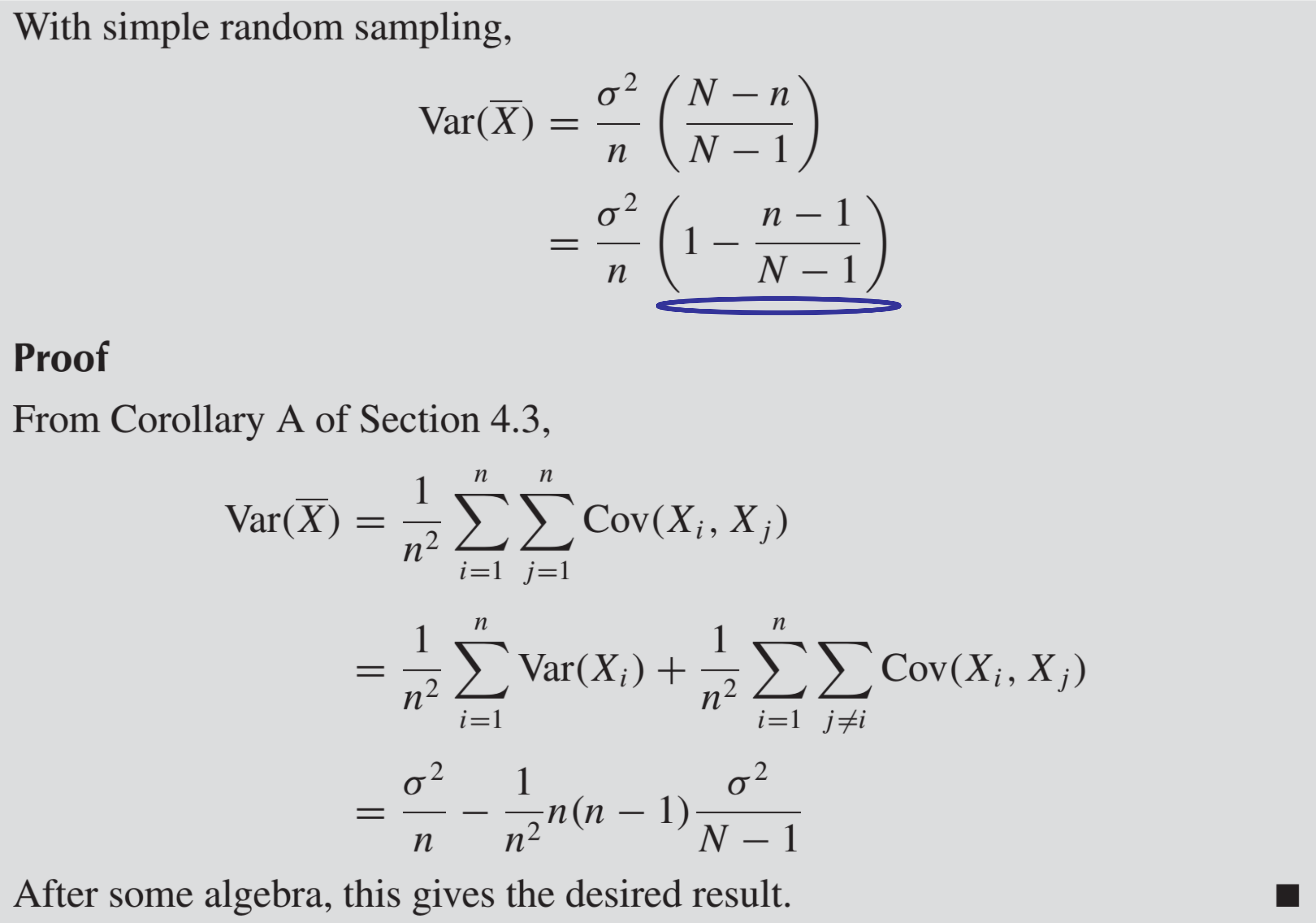
Sample Mean: Its variance

Using



$$\mathrm{Cov}(X_i, X_j) = -\sigma^2/(N-1) \qquad \text{if } i \neq j$$

we now have the following theorem:



With simple random sampling,

$$\begin{aligned}\mathrm{Var}(\overline{X}) &= \frac{\sigma^2}{n}\left(\frac{N-n}{N-1}\right)\\ &= \frac{\sigma^2}{n}\left(1-\frac{n-1}{N-1}\right)\end{aligned}$$

Proof

From Corollary A of Section 4.3,

$$\begin{aligned}\mathrm{Var}(\overline{X}) &= \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\mathrm{Cov}(X_i, X_j)\\ &= \frac{1}{n^2}\sum_{i=1}^{n}\mathrm{Var}(X_i) + \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j\neq i}\mathrm{Cov}(X_i, X_j)\\ &= \frac{\sigma^2}{n} - \frac{1}{n^2}n(n-1)\frac{\sigma^2}{N-1}\end{aligned}$$

After some algebra, this gives the desired result. ■

Sample Mean: Its variance

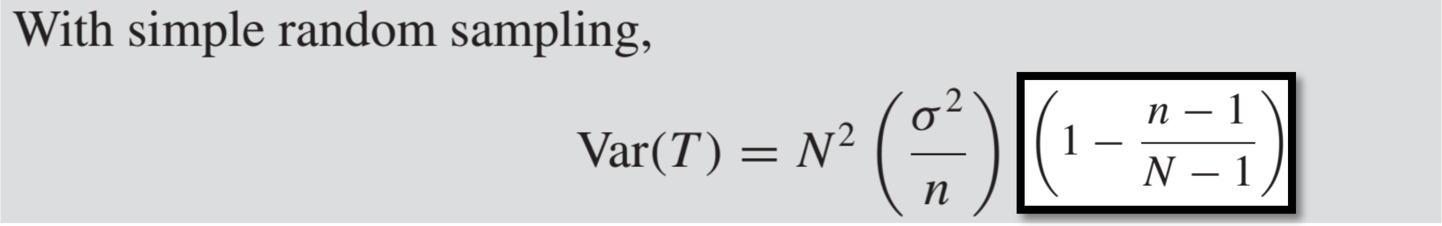
Variance of the sample mean in sampling without replacement differs from that in sampling with replacement by the **finite population correction**:

$$\boxed{\left(1-\frac{n-1}{N-1}\right)}$$

Frequently the sampling fraction *n/N* is very small, and the standard error of $\overline{X}$ is

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

Apart from the usually small finite population correction, the spread of the sampling distribution and thus the precision of *X-bar* are determined by the sample size (*n*) and not by the population size (*N*).



With simple random sampling,

$$\operatorname{Var}(T) = N^2\left(\frac{\sigma^2}{n}\right)\boxed{\left(1-\frac{n-1}{N-1}\right)}$$

Sample Mean: Its variance

Example. Hospitals again: Sampling without replacement, sample size=32,

$$\sigma_{\overline{X}} = \frac{\sigma}{\sqrt{n}}\sqrt{1-\frac{n-1}{N-1}} = \frac{589.7}{\sqrt{32}}\sqrt{1-\frac{31}{392}}$$

$$= 104.2 \times 0.96 = 100.0$$

Sample Mean: Its variance

Example. Hospitals again: Sampling without replacement, sample size=32,

$$\sigma_{\overline{X}} = \frac{\sigma}{\sqrt{n}}\sqrt{1-\frac{n-1}{N-1}} = \frac{589.7}{\sqrt{32}}\sqrt{1-\frac{31}{392}}$$
$$= 104.2 \times 0.96 = 100.0$$

Example. Still hospitals. Estimating a proportion.

A proportion p = .654 had fewer than 1000 discharges. If this proportion were estimated from a sample as the sample proportion $\hat{p}$, the standard error of $\hat{p}$

$$\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}\sqrt{1-\frac{n-1}{N-1}}$$

For n=32,

$$\sigma_{\hat{p}} = \sqrt{\frac{0.654 \times 0.346}{32}}\sqrt{1-\frac{31}{392}} = 0.08$$

Estimation of the population variance

A sample survey is used to estimate population parameters and assess and quantify the variability of the estimates. We saw how the standard error of an estimate may be determined from the sample size and the population variance.

So far so good… No, the population variance is unknown! Estimate it?

This looks natural:

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

But actually, this estimate is biased.

Estimation of the population variance

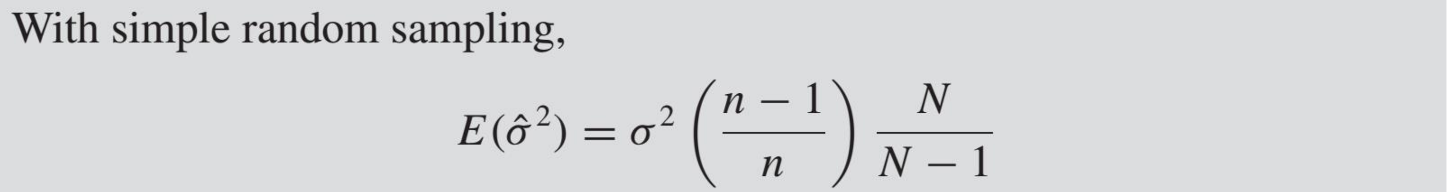
A sample survey is used to estimate population parameters and assess and quantify the variability of the estimates. We saw how the standard error of an estimate may be determined from the sample size and the population variance.

So far so good… No, the population variance is unknown! Estimate it?

This looks natural:

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

But actually, this estimate is biased.



With simple random sampling,

$$E(\hat{\sigma}^2) = \sigma^2\left(\frac{n-1}{n}\right)\frac{N}{N-1}$$

With simple random sampling,

$$E(\hat{\sigma}^2) = \sigma^2 \left(\frac{n-1}{n}\right) \frac{N}{N-1}$$

Proof

Expanding the square and proceeding as in the identity for the population variance in Section 7.2, we find

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \overline{X}^2$$

Thus,

$$E(\hat{\sigma}^2) = \frac{1}{n}\sum_{i=1}^{n} E\left(X_i^2\right) - E(\overline{X}^2)$$

Now, we know that

$$\begin{aligned} E\left(X_i^2\right) &= \mathrm{Var}(X_i) + [E(X_i)]^2 \\ &= \sigma^2 + \mu^2 \end{aligned}$$

Similarly, from Theorems A and B of Section 7.3.1,

$$\begin{aligned} E(\overline{X}^2) &= \mathrm{Var}(\overline{X}) + [E(\overline{X})]^2 \\ &= \frac{\sigma^2}{n}\left(1 - \frac{n-1}{N-1}\right) + \mu^2 \end{aligned}$$

Substituting these expressions for $E(X_i^2)$ and $E(\overline{X}^2)$ in the preceding equation for $E(\hat{\sigma}^2)$ gives the desired result. ■

Estimation of the population variance

For *N*>*n*,

$$\frac{n-1}{n}\frac{N}{N-1} < 1$$

so that $E(\hat{\sigma}^2) < \sigma^2$; $\boxed{\hat{\sigma}^2 \text{ thus tends to underestimate } \sigma^2.}$

Correcting with this factor, an unbiased estimate is

$$\frac{1}{n-1}\left(1-\frac{1}{N}\right)\sum_{i=1}^{n}(X_i-\overline{X})^2.$$

Estimation of the population variance

For *N*>*n*,

$$\frac{n-1}{n}\frac{N}{N-1} < 1$$

so that $E(\hat{\sigma}^2) < \sigma^2$; $\boxed{\hat{\sigma}^2 \text{ thus tends to underestimate } \sigma^2.}$

Correcting with this factor, an unbiased estimate is

$$\boxed{\frac{1}{n-1}\left(1-\frac{1}{N}\right)\sum_{i=1}^{n}(X_i-\overline{X})^2.}$$

Since

$$\mathrm{Var}\,(\overline{X}) = \frac{\sigma^2}{n}\left(\frac{N-n}{N-1}\right)$$

we know:

An unbiased estimate of $\mathrm{Var}(\overline{X})$ is

$$\begin{aligned} s_{\overline{X}}^2 &= \frac{\hat{\sigma}^2}{n}\left(\frac{n}{n-1}\right)\left(\frac{N-1}{N}\right)\left(\frac{N-n}{N-1}\right) \\ &= \frac{s^2}{n}\left(1-\frac{n}{N}\right) \end{aligned}$$

where

Sample variance: $\boxed{s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i-\overline{X})^2}$

Estimation of the population variance

Similarly, an unbiased estimate of the variance of *T*, the estimator of the population total is

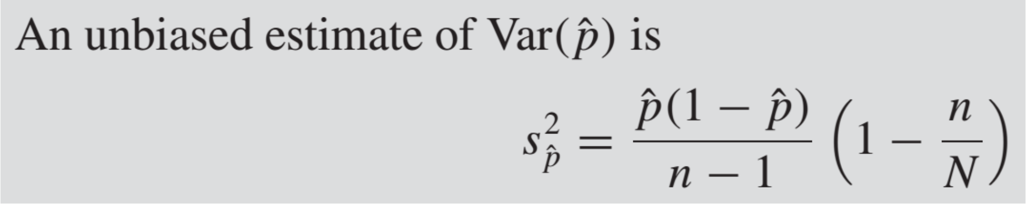
$$s_T^2 = N^2 s_{\overline{X}}^2$$

For the dichotomous case, in which each X_i is 0 or 1,

$$\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \overline{X}^2 = \hat{p}(1-\hat{p})$$

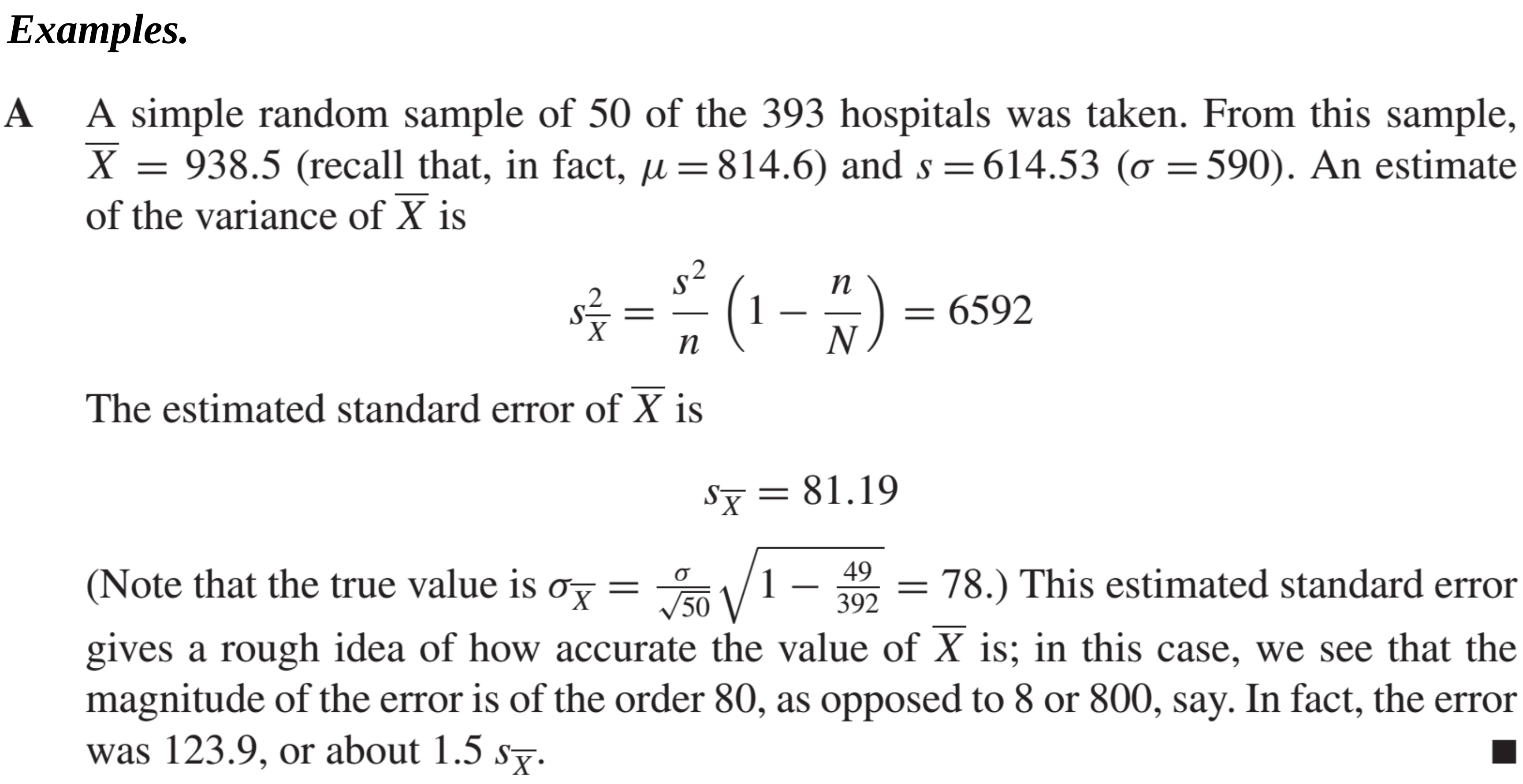
$$s^2 = \frac{n}{n-1}\hat{p}(1-\hat{p})$$

Therefore, we have



An unbiased estimate of Var$(\hat{p})$ is

$$s_{\hat{p}}^2 = \frac{\hat{p}(1-\hat{p})}{n-1}\left(1 - \frac{n}{N}\right)$$



Examples.

A A simple random sample of 50 of the 393 hospitals was taken. From this sample, $\overline{X} = 938.5$ (recall that, in fact, $\mu = 814.6$) and $s = 614.53$ ($\sigma = 590$). An estimate of the variance of $\overline{X}$ is

$$s_{\overline{X}}^2 = \frac{s^2}{n}\left(1 - \frac{n}{N}\right) = 6592$$

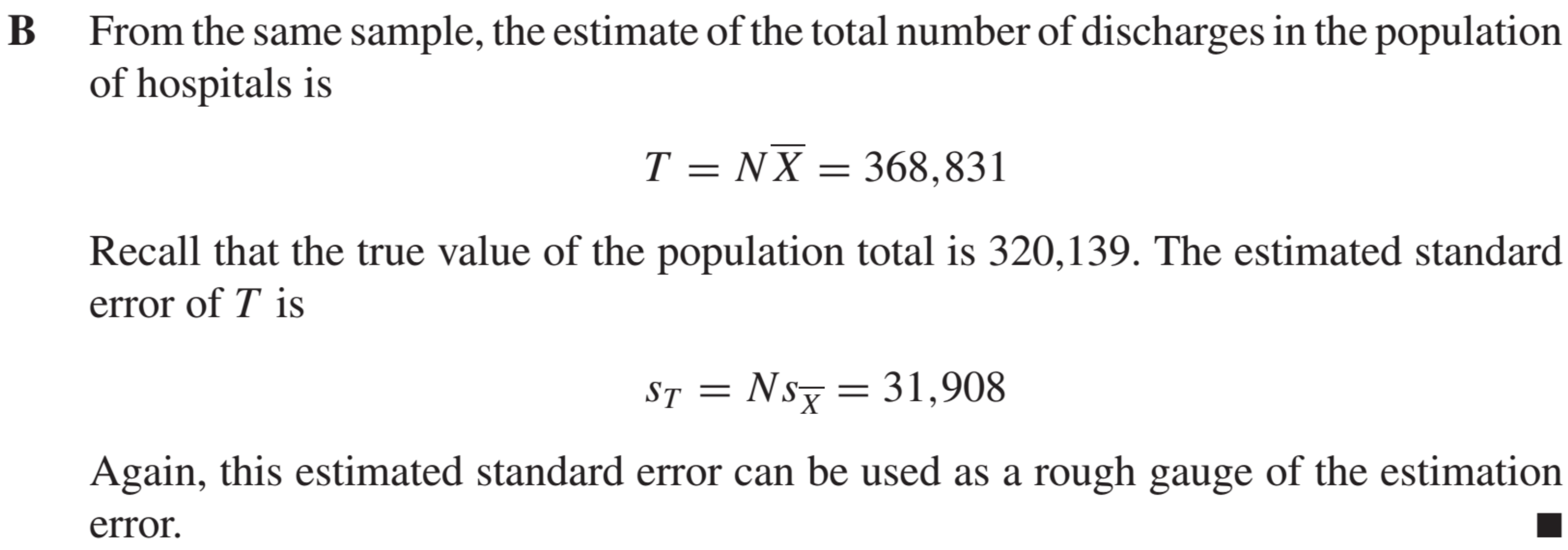
The estimated standard error of $\overline{X}$ is

$$s_{\overline{X}} = 81.19$$

(Note that the true value is $\sigma_{\overline{X}} = \frac{\sigma}{\sqrt{50}}\sqrt{1 - \frac{49}{392}} = 78$.) This estimated standard error gives a rough idea of how accurate the value of $\overline{X}$ is; in this case, we see that the magnitude of the error is of the order 80, as opposed to 8 or 800, say. In fact, the error was 123.9, or about 1.5 $s_{\overline{X}}$. ■

Mean=814.6,
Total=320,138,
Variance=347,766
Standard deviation=589.7

Examples.



B From the same sample, the estimate of the total number of discharges in the population of hospitals is

$$T = N\overline{X} = 368{,}831$$

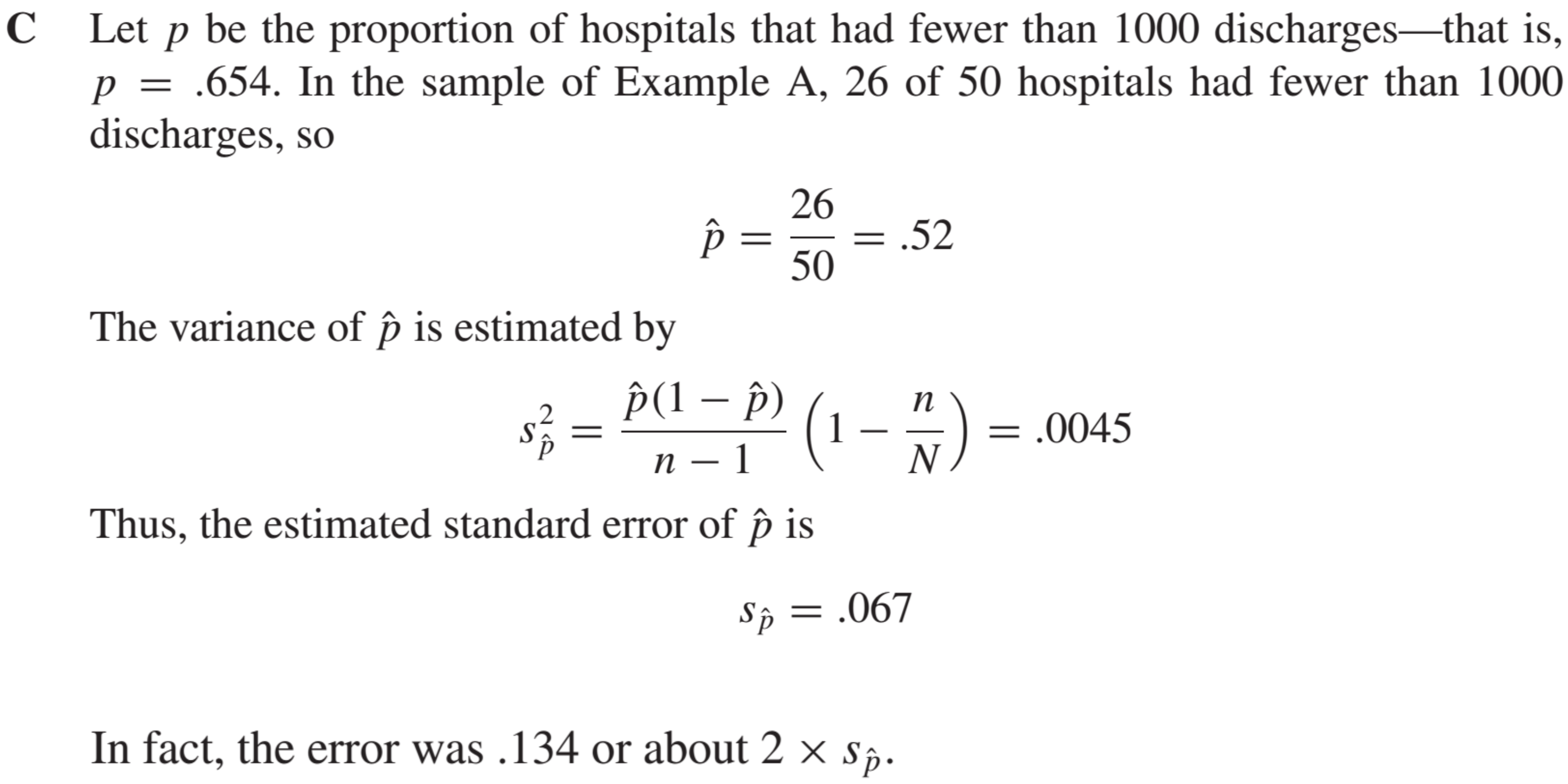
Recall that the true value of the population total is 320,139. The estimated standard error of T is

$$s_T = Ns_{\overline{X}} = 31{,}908$$

Again, this estimated standard error can be used as a rough gauge of the estimation error. ■

Mean=814.6,
Total=320,138,
Variance=347,766
Standard deviation=589.7

Examples.



C Let p be the proportion of hospitals that had fewer than 1000 discharges—that is, $p = .654$. In the sample of Example A, 26 of 50 hospitals had fewer than 1000 discharges, so

$$\hat{p} = \frac{26}{50} = .52$$

The variance of $\hat{p}$ is estimated by

$$s_{\hat{p}}^2 = \frac{\hat{p}(1-\hat{p})}{n-1}\left(1 - \frac{n}{N}\right) = .0045$$

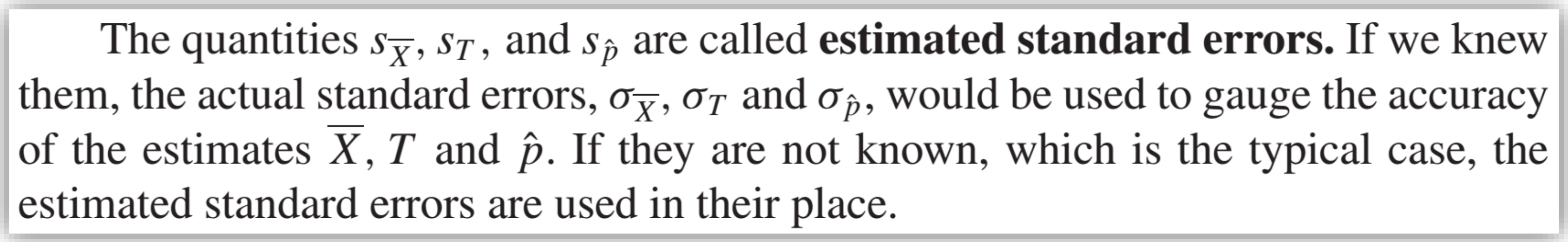
Thus, the estimated standard error of $\hat{p}$ is

$$s_{\hat{p}} = .067$$

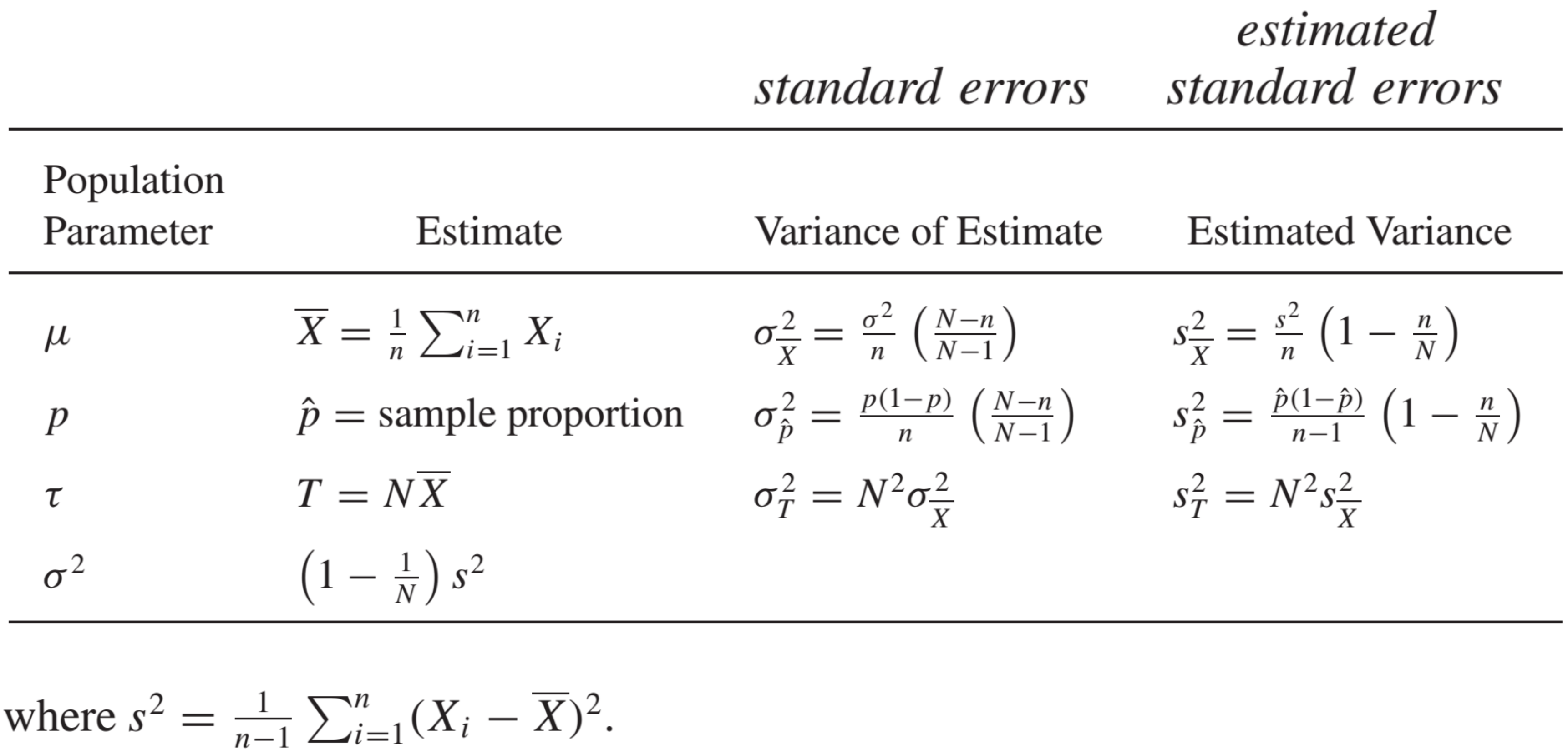
In fact, the error was .134 or about 2 × $s_{\hat{p}}$.

Mean=814.6,
Total=320,138,
Variance=347,766
Standard deviation=589.7

Everything together!



The quantities $s_{\overline{X}}$, s_T, and $s_{\hat{p}}$ are called **estimated standard errors.** If we knew them, the actual standard errors, $\sigma_{\overline{X}}$, σ_T and $\sigma_{\hat{p}}$, would be used to gauge the accuracy of the estimates $\overline{X}$, T and $\hat{p}$. If they are not known, which is the typical case, the estimated standard errors are used in their place.



		standard errors	*estimated* *standard errors*
Population Parameter	Estimate	Variance of Estimate	Estimated Variance
μ	$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$	$\sigma_{\overline{X}}^2 = \frac{\sigma^2}{n}\left(\frac{N-n}{N-1}\right)$	$s_{\overline{X}}^2 = \frac{s^2}{n}\left(1 - \frac{n}{N}\right)$
p	$\hat{p} =$ sample proportion	$\sigma_{\hat{p}}^2 = \frac{p(1-p)}{n}\left(\frac{N-n}{N-1}\right)$	$s_{\hat{p}}^2 = \frac{\hat{p}(1-\hat{p})}{n-1}\left(1 - \frac{n}{N}\right)$
τ	$T = N\overline{X}$	$\sigma_T^2 = N^2\sigma_{\overline{X}}^2$	$s_T^2 = N^2 s_{\overline{X}}^2$
σ^2	$\left(1 - \frac{1}{N}\right) s^2$		

where $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$.

Normal approx. to sampling distribution of sample mean

Ideally, we desire to know the sampling distribution, which can tell us everything we hope to know about the accuracy of the estimate. However, this is not feasible without knowledge of the population itself.

But we notice that the simulation of sample mean is roughly Gaussian. – We recall that CLT is so overwhelming that *n* does not have to be that large... *We can use CLT to deduce an approximation to the sampling distribution, and find probabilistic bounds for the estimation error using this approximation.*

Consider a sequence of independent and identically distributed (i.i.d.) random variables, $X_1, X_2, \ldots$ having the common mean and variance μ and σ^2. The sample mean is

$$\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i \qquad E(\overline{X}_n) = \mu \qquad \operatorname{Var}(\overline{X}_n) = \frac{\sigma^2}{n}$$

CLT says that, for a fixed number *z*,

$$P\left(\frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}} \leq z\right) \to \Phi(z) \qquad \text{as } n \to \infty$$

$$\boxed{P\left(\frac{\overline{X}_n - \mu}{\sigma_{\overline{X}_n}} \leq z\right) \to \Phi(z)}$$

"Standardized"

Cdf of standard normal distribution

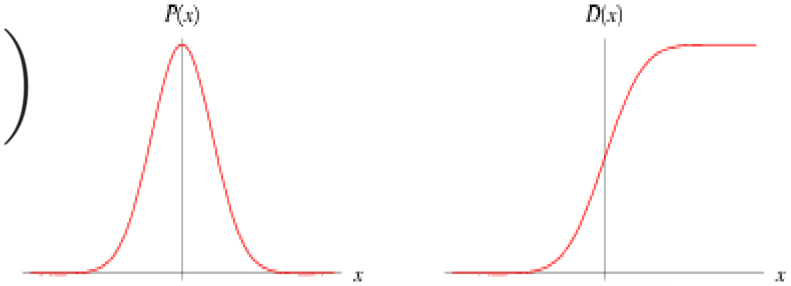
Normal approx. to sampling distribution of sample mean

However, the context here is not exactly like that of CLT!

- Sampling without replacement, X_i are **not independent** of each other
- It makes no sense to have *n* tend to infinity while *N* remains fixed…
- But ***we have many CLTs***, other CLTs have been proved appropriate to the sampling context: if *n* is large, though still small relative to *N*, then the mean of a simple random sample is approximately normally distributed

Probability that the error made in estimating μ by *X-bar* is less than some constant δ:

$$\begin{aligned} P(|\overline{X} - \mu| \leq \delta) &= P(-\delta \leq \overline{X} - \mu \leq \delta) \\ &= P\left(-\frac{\delta}{\sigma_{\overline{X}}} \leq \frac{\overline{X} - \mu}{\sigma_{\overline{X}}} \leq \frac{\delta}{\sigma_{\overline{X}}}\right) \\ &\approx \Phi\left(\frac{\delta}{\sigma_{\overline{X}}}\right) - \Phi\left(-\frac{\delta}{\sigma_{\overline{X}}}\right) \\ &= 2\Phi\left(\frac{\delta}{\sigma_{\overline{X}}}\right) - 1 \end{aligned}$$



$$\boxed{\Phi(-z) = 1 - \Phi(z)}$$

Normal approx. to sampling distribution of sample mean

Example A: The population of 393 Hospitals again. Sample size *n*=64, using the finite population correction, sample mean’s standard deviation is

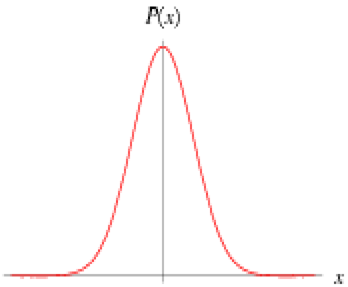
$$\sigma_{\overline{X}} = \frac{\sigma}{\sqrt{n}}\sqrt{1 - \frac{n-1}{N-1}}$$
$$= \frac{589.7}{8}\sqrt{1 - \frac{63}{392}} = 67.5$$

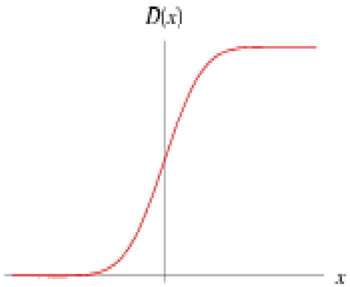
Mean=814.6,
Total=320,138,
Standard deviation=589.7

Use CLT to approximate the probability that the sample mean differs from the population mean by more than 100 is ~14% (16.4% is a real number):

$$P(|\overline{X} - \mu| > 100) \approx 2P(\overline{X} - \mu > 100)$$

$$P(\overline{X} - \mu > 100) = 1 - P(\overline{X} - \mu < 100)$$
$$= 1 - P\left(\frac{\overline{X} - \mu}{\sigma_{\overline{X}}} < \frac{100}{\sigma_{\overline{X}}}\right)$$
$$\approx 1 - \Phi\left(\frac{100}{67.5}\right)$$
$$= .069$$





Normal approx. to sampling distribution of sample mean

Example B: For a sample of size 50, the standard error of the sample mean number of discharges is

$$\sigma_{\overline{X}} = 78$$

$\overline{X} = 938.5$, so $\overline{X} - \mu = 123.9.$

What is the probability of an error this large or larger?

$$\begin{aligned} P(|\overline{X} - \mu| \geq 123.9) &= 1 - P(|\overline{X} - \mu| < 123.9) \\ &\approx 1 - \left[2\Phi\left(\frac{123.9}{78}\right) - 1\right] \\ &= 2 - 2\Phi(1.59) \\ &= .11 \end{aligned}$$

Mean=814.6,
Total=320,138,
Standard deviation=589.7

Normal approx. to sampling distribution of sample mean

Example C: We find from the sample of size 50 an estimate $\hat{p} = .52$ of the proportion of hospitals that discharged fewer than 1000 patients; in fact, the actual proportion in the population is .65. Thus,

$$|\hat{p} - p\,| = .13.$$

What is the probability that an estimate will be off by an amount this large or larger?

$$\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}\sqrt{1 - \frac{n-1}{N-1}}$$
$$= .068 \times .94 = .064$$

Mean=814.6,
Total=320,138,
Standard deviation=589.7

$$\begin{aligned} P(|p - \hat{p}| > .13) &= 1 - P(|p - \hat{p}| \leq .13) \\ &= 1 - P\left(\frac{|p-\hat{p}|}{\sigma_{\hat{p}}} \leq \frac{.13}{\sigma_{\hat{p}}}\right) \\ &\approx 2[1 - \Phi(2.03)] = .04 \end{aligned}$$

Confidence interval

A confidence interval for a population parameter, θ, is a random interval, calculated from the sample, that contains θ with some specified probability.

A 95% confidence interval for μ is a random interval that contains μ with probability .95; if we were to take many random samples and form a confidence interval from each one, about 95% of these intervals would contain μ.

If the coverage probability=1−α, the interval is called a 100(1−α)% confidence interval.

For $0 \le \alpha \le 1$, let $z(\alpha)$ be the number such that the area under the standard normal pdf to the right of $z(\alpha)$ is α. (上**α**分位点)

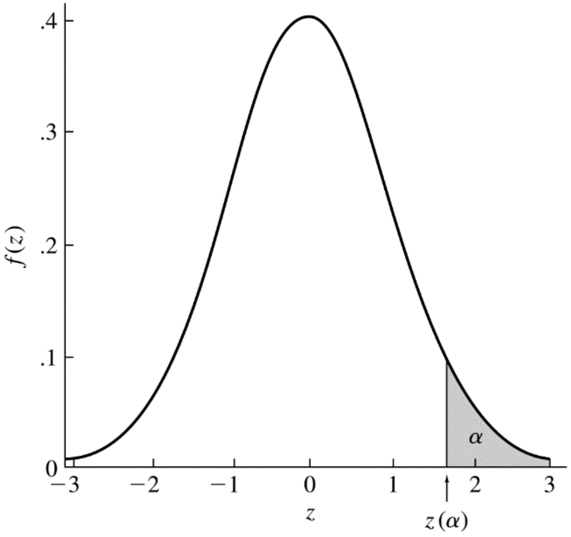
$z(1-\alpha)=-z(\alpha)$

If $Z\sim N(0,1)$, $P(-z(\alpha/2) \le Z \le z(\alpha/2))=1-\alpha$.

CLT: $$P\left(-z(\alpha/2) \le \frac{\overline{X}-\mu}{\sigma_{\overline{X}}} \le z(\alpha/2)\right) \approx 1-\alpha$$

$$\boxed{P(\overline{X} - z(\alpha/2)\sigma_{\overline{X}} \le \mu \le \overline{X} + z(\alpha/2)\sigma_{\overline{X}}) \approx 1-\alpha}$$

区间为平均值±数倍σ形式



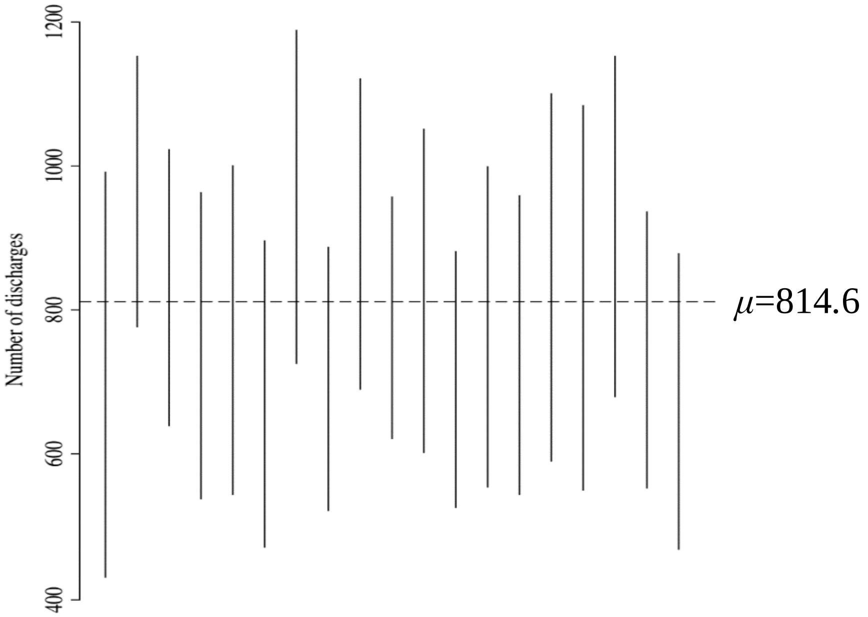
Confidence interval

$$\boxed{P(\overline{X} - z(\alpha/2)\sigma_{\overline{X}} \le \mu \le \overline{X} + z(\alpha/2)\sigma_{\overline{X}}) \approx 1 - \alpha}$$

The probability that μ lies in the interval $\overline{X} \pm z(\alpha/2)\sigma_{\overline{X}}$ is approximately $1-\alpha$.

- Note and understand that this interval is random
- In practice, α is assigned small values, 0.1, 0.05, 0.01, for large coverage probability
- Population variance is typically unknown, $\sigma_{\overline{X}}$ is substituted by $s_{\overline{X}}$ if sample is large
- How large is large? Rule of thumb: *n* greater than 25 or 30 is usually adequate

Example. 20 samples each of size *n*=25 are drawn from the population of hospital discharges. From each samples, an approximate 95% confidence interval for μ, the mean number of discharges, was computed. Among the 20 confidence intervals (vertical lines), ~1 out of 20 does not include μ.



Confidence interval

Example D. A particular area contains 8000 condominium (公寓) units. In a survey of the occupants, a simple random sample of size 100 yields the information that the average number of motor vehicles per unit is 1.6, with a sample standard deviation of 0.8. The estimated standard error of *X-bar* is thus

$$
\begin{aligned}
s_{\overline{X}} &= \frac{s}{\sqrt{n}}\sqrt{1-\frac{n}{N}} \\
&= \frac{.8}{10}\sqrt{1-\frac{100}{8000}} \\
&= .08
\end{aligned}
$$

Note that the finite population correction makes almost no difference. Since $z(.025)=1.96$, a 95% confidence interval for the population average is $X \pm 1.96\, s_{\overline{X}}$, or (1.44,1.76).

An estimate of the total number of motor vehicles is $T=8000\times 1.6=12{,}800$. The estimated standard error of T is $s_T = N s_{\overline{X}} = 640$. A 95% confidence interval for the total # of motor vehicles is $T \pm 1.96 s_T$, or (11,546, 14,054).

Confidence interval

Example D’. In the same survey, 12% of the respondents said they planned to sell their condos within the next year; $\hat{p}$ = .12 is an estimate of the population proportion p. The estimated standard error is

$$s_{\hat{p}} = \sqrt{\frac{\hat{p}(1-\hat{p})}{n-1}}\sqrt{1-\frac{100}{8000}} = .03$$

A 95% confidence interval for p is $\hat{p} \pm 1.96 s_{\hat{p}}$, or (.06, .18).

The total number of owners planning to sell is estimated as $T = N\hat{p} = 960$. The estimated standard error of T is $s_T = N s_{\hat{p}} = 240$. A 95% confidence interval for the number in the population planning to sell is $T \pm 1.96 s_T$, or (490, 1430).

Confidence interval

Example D’. In the same survey, 12% of the respondents said they planned to sell their condos within the next year; $\hat{p}$ = .12 is an estimate of the population proportion *p*. The estimated standard error is

$$s_{\hat{p}} = \sqrt{\frac{\hat{p}(1-\hat{p})}{n-1}}\sqrt{1-\frac{100}{8000}} = .03$$

A 95% confidence interval for p is $\hat{p} \pm 1.96s_{\hat{p}}$, or (.06, .18).

The total number of owners planning to sell is estimated as $T = N\,\hat{p} = 960$. The estimated standard error of T is $s_T = Ns_{\hat{p}} = 240$. A 95% confidence interval for the number in the population planning to sell is $T \pm 1.96s_T$, or (490, 1430).

Proper interpretation of this interval is a little subtle. We cannot state that the probability is 0.95 and that the # of owners planning to sell is between 490 and 1430 (that # is either in this interval or not).

95% of intervals formed in this way will contain the true number in the long run.

In the long run, 95% of those intervals will contain the true number of discharges, but in the figure any particular interval either does or doesn’t contain the true number (cf. figure on slide pp. 43).

Confidence interval

The width of a confidence interval is determined by the sample size n and the population standard deviation σ.
If σ is known approximately, perhaps from earlier samples of the population, n can be chosen so as to obtain a confidence interval close to some desired length.
-- An important aspect of planning the design of a sample survey.

Example E.

The interval for the total # of owners planning to sell in Example D might be considered too wide for practical purposes; reducing its width would require a larger sample size.

Suppose that an interval with a half-width of 200 is desired. Neglecting the finite population correction, the half-width is

$$1.96 s_T = 1.96 N \sqrt{\frac{\hat{p}(1-\hat{p})}{n-1}} = \frac{5095}{\sqrt{n-1}}$$

Set it to 200 yields *n*=650 as the necessary sample size.

Summary

- Fundamental result: the sampling distribution of the sample mean is approximately Gaussian.

- This approximation can be used to quantify the error committed in estimating the population mean by the sample mean, giving us a good understanding of the accuracy of estimates produced by a simple random sample.

- confidence interval, a random interval that contains a population parameter with a specified probability, and thus provides an assessment of the accuracy of the corresponding estimate of that parameter.

- We have seen in our examples that the width of the confidence interval is a multiple of the estimated standard deviation of the estimate a confidence interval for *μ* is $\overline{X} \pm k s_{\overline{X}}$, where the constant *k* depends on the coverage probability of the interval.

Estimation of a ratio

Estimation of a ratio

Suppose that for each member of a population, two values, *x* and *y*, may be measured. The ratio of interest is

$$r = \frac{\sum_{i=1}^{N} y_i}{\sum_{i=1}^{N} x_i} = \frac{\mu_y}{\mu_x}$$

- If *y* is the # of unemployed males aged 20-30 in a household and *x* is the # of males aged 20-30 in a household, then *r* is the proportion of unemployed males aged 20-30.
- If *y* is weekly food expenditure and *x* is # of inhabitants, then *r* is weekly food cost per inhabitant.
- If *y* is the # of motor vehicles and *x* is the number of inhabitants of driving age, then *r* is the # of motor vehicles per inhabitant of driving age.
- In a survey of farms, *y* might be the acres of wheat planted and *x* the total acreage.
- In an inventory audit（库存审计）, *y* might be the audited value of an item and *x* the book value （账面值）.

Estimation of a ratio

First of all, note

$$r \neq \frac{1}{N}\sum_{i=1}^{N}\frac{y_i}{x_i}$$

The natural estimate of r is

$$R = \overline{Y}/\overline{X}.$$

We wish to derive expressions for E(R), Var(R), but for this non-linear function we need the approximation methods we developed before.

To calculated the approx. variance of R, we need

$$\mathrm{Var}(\overline{X}),\ \mathrm{Var}(\overline{Y}),\ \text{and}\ \mathrm{Cov}(\overline{X}, \overline{Y}).$$

For the last quantity, we define the **population covariance** of *x* and *y* to be

$$\sigma_{xy} = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu_x)(y_i - \mu_y)$$

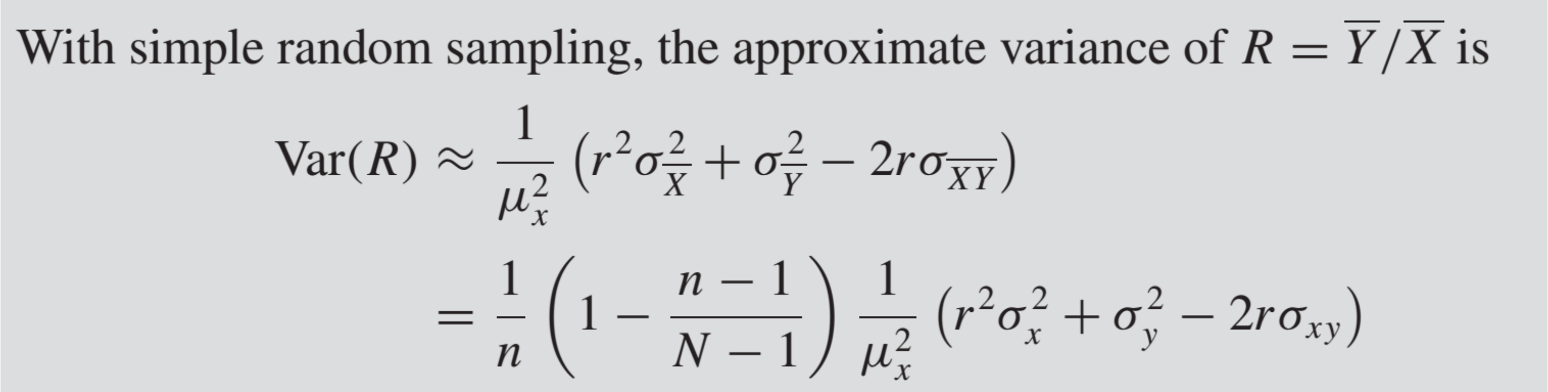
Then it can be shown that

$$\mathrm{Cov}(\overline{X}, \overline{Y}) = \frac{\sigma_{xy}}{n}\left(1 - \frac{n-1}{N-1}\right)$$

Analogous to

$$\mathrm{Cov}(X_i, X_j) = -\sigma^2/(N-1) \qquad \text{if } i \neq j$$

Estimation of a ratio



With simple random sampling, the approximate variance of $R = \overline{Y}/\overline{X}$ is

$$\begin{aligned}\mathrm{Var}(R) &\approx \frac{1}{\mu_x^2}\left(r^2\sigma_{\overline{X}}^2 + \sigma_{\overline{Y}}^2 - 2r\sigma_{\overline{XY}}\right)\\ &= \frac{1}{n}\left(1 - \frac{n-1}{N-1}\right)\frac{1}{\mu_x^2}\left(r^2\sigma_x^2 + \sigma_y^2 - 2r\sigma_{xy}\right)\end{aligned}$$

The **population correlation coefficient** is defined as

$$\rho = \frac{\sigma_{xy}}{\sigma_x\sigma_y}$$

$$\mathrm{Var}(R) \approx \frac{1}{n}\left(1 - \frac{n-1}{N-1}\right)\frac{1}{\mu_x^2}\left(r^2\sigma_x^2 + \sigma_y^2 - 2r\rho\sigma_x\sigma_y\right)$$

Problem set #3

1. Utilizing the powerful mgfs, prove an important property of the χ^2 distribution: if U and V are independent and $U \sim \chi_n{}^2$ and $V \sim \chi_n{}^2$, then $U+V \sim \chi_{m+n}{}^2$.
2. Prove that the mgf of $\chi_n{}^2$ distribution is $M(t)=(1-2t)^{-n/2}$, and thus the mean is n and the variance is $2n$. Plot the **reduced** χ^2 distribution for n=2, 4, 6, *you need some intuition about this distribution, which prevails over the whole astronomy*.
3. Find expressions for the approximate mean and variance of $Y = \ln x$.
4. In Lecture 5 we examined the accuracy of our approximations for the case of $g(x)=x^{1/2}$. Do the same thing for $g(x)=x^{1/3}$, i.e. suppose X is uniform on [0, 1] and on [1, 2], compare the approximated mean and variance to their corresponding exact values, and show on which interval the approximation works better.
5. Verify that the Gaussian integral in our example is indeed .3417, when the Monte Carlo method is employed. Change the integration limits to [0, 5] and see if it is close to ½.
6. In problem set #1, you learned how to create exponential random variables. Based on that, reproduce the four panels of the figure on the last page (overplotting the best-fit Gaussian profile is optional). Note: you need to do averaging here instead of summing, though they are *virtually* the same.