The Basic Practice of Statistics
Ninth Edition

David S. Moore William I. Notz

Chapter 15
Sampling Distributions

Lecture Slides

In Chapter 15, we cover …

- Parameters and statistics
- Statistical estimation and the law of large numbers
- Sampling distributions
- The sampling distribution of $\bar{x}$
- The central limit theorem
- Sampling distributions and statistical significance

Parameters and statistics

- As we begin to use sample data to draw conclusions about a wider population, we must be clear about whether a number describes a sample or a population.

- A **parameter** is a number that describes the population. In practice, the value of a parameter is not known because we can rarely examine the entire population.

- A **statistic** is a number that can be computed from the sample data without making use of any unknown parameters. In practice, we often use a statistic to estimate an unknown parameter.

Example 15.1

- The mean income of the sample of 128,579 households included in the 2018 Current Population Survey was $\bar{x} = \$90{,}021$.
- The number $90,021 is a *statistic* because it describes this one Current Population Survey sample.
- The population that the poll wants to draw conclusions about is all 128 million U.S. households.
- The *parameter* of interest is the mean income of all these households. We don't know the value of this parameter.

Parameters and statistics

- Remember ***p*** and ***s***: ***p***arameters come from ***p***opulations and ***s***tatistics come from ***s***amples.

- We write μ (the Greek letter mu) for the **mean of the population** and σ (the Greek letter sigma) for the **standard deviation of the population**. We write $\bar{x}$ (“x-bar”) for the **mean of the sample** and ***s*** for the **standard deviation of the sample**.

Statistical estimation

- The process of **statistical inference** involves using information from a sample to draw conclusions about a wider population.
- Different random samples yield different statistics. We need to be able to describe the **sampling distribution** of possible statistic values in order to perform statistical inference.
- We can think of a statistic as a **random variable** because it takes numerical values that describe the outcomes of the random sampling process. Therefore, we can examine its probability distribution using concepts we learned in earlier chapters.

The law of large numbers

- If $\bar{x}$ is rarely exactly right and varies from sample to sample, why is it nonetheless a reasonable estimate of the population mean μ?
- Here is one answer: If we keep taking larger and larger samples, the statistic $\bar{x}$ is guaranteed to get closer and closer to the parameter μ.

LAW OF LARGE NUMBERS

- Draw observations at random from any population with finite mean μ. As the number of observations drawn increases, the mean $\bar{x}$ of the observed values tends to get closer and closer to the mean μ of the population.

Example law of large numbers (1 of 3)

Example week 12

In [1]:
```
import numpy as np
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
```

np.random.chisquare: https://numpy.org/doc/stable/reference/random/generated/numpy.random.chisquare.html

In [2]:
```
# Let us simulate data from a population of 100K individuals.
# Suppose the data correspond to the number of hours spent by each individual
# on a certain streaming service in October 2021
# the true population mean is 20 hours
np.random.seed(0)
hours = np.random.chisquare(df = 20, size = 100000)
df_hours = pd.DataFrame({'hours': hours})
```

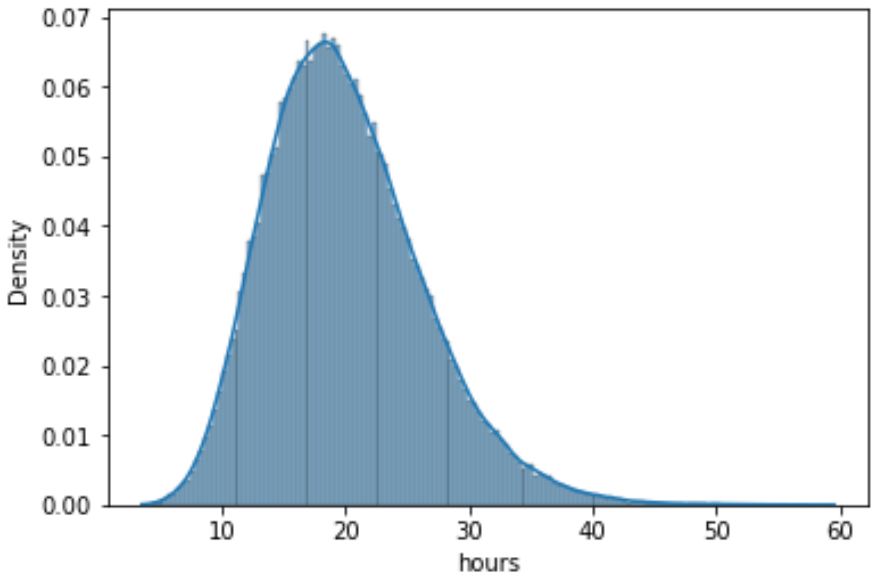
In [3]:
```
df_hours.head()
```

Out[3]:

	hours
0	32.508030
1	21.929889
2	33.426648
3	13.870850
4	18.698569

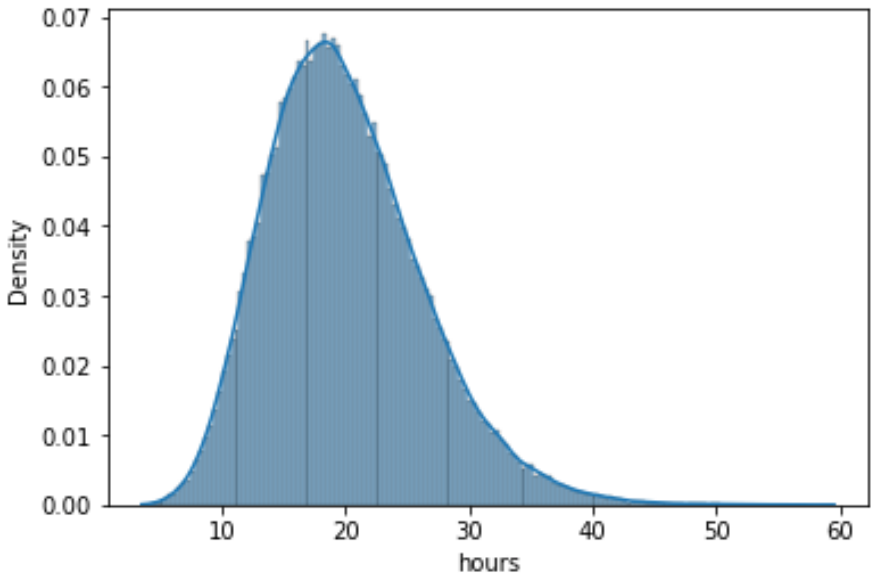
Example using Python (2 of 3)

In [4]:
```
# Population distribution is not Normal!
# It is a Chi-square distribution with 20 degrees of freedom, a right-skewed distribution
sns.histplot(df_hours['hours'], stat = 'density', kde = True)
plt.show()
```



Example law of large numbers (1 of 3)

In [4]:
```
# Population distribution is not Normal!
# It is a Chi-square distribution with 20 degrees of freedom, a right-skewed distribution
sns.histplot(df_hours['hours'], stat = 'density', kde = True)
plt.show()
```



Example law of large numbers (1 of 3)

In [5]:
```
# Law of large numbers
# As we increase the sample size, the mean estimate gets closer and closer to the true population mean
sample_size = [5,10,50,100,5000]
np.random.seed(5)
for n in sample_size:
    sample = np.random.choice(df_hours['hours'], size = n, replace=False)
    sample_mean = np.mean(sample)
    print('sample size = ',n)
    print('sample mean = ',sample_mean)
```

```
sample size =  5
sample mean =  22.099435416069532
sample size =  10
sample mean =  17.76744417315056
sample size =  50
sample mean =  19.597336694207254
sample size =  100
sample mean =  20.535484416927055
sample size =  5000
sample mean =  20.055075192756277
```

As we increase the sample size n, $\bar{x}$ gets closer and closer to μ

Sampling distributions

- The law of large numbers assures us that if we measure enough subjects, the statistic $\bar{x}$ will eventually get very close to the unknown parameter μ.

- If we took every one of the possible samples of a certain size, calculated the sample mean for each, and graphed all of those values, we'd have a sampling distribution.

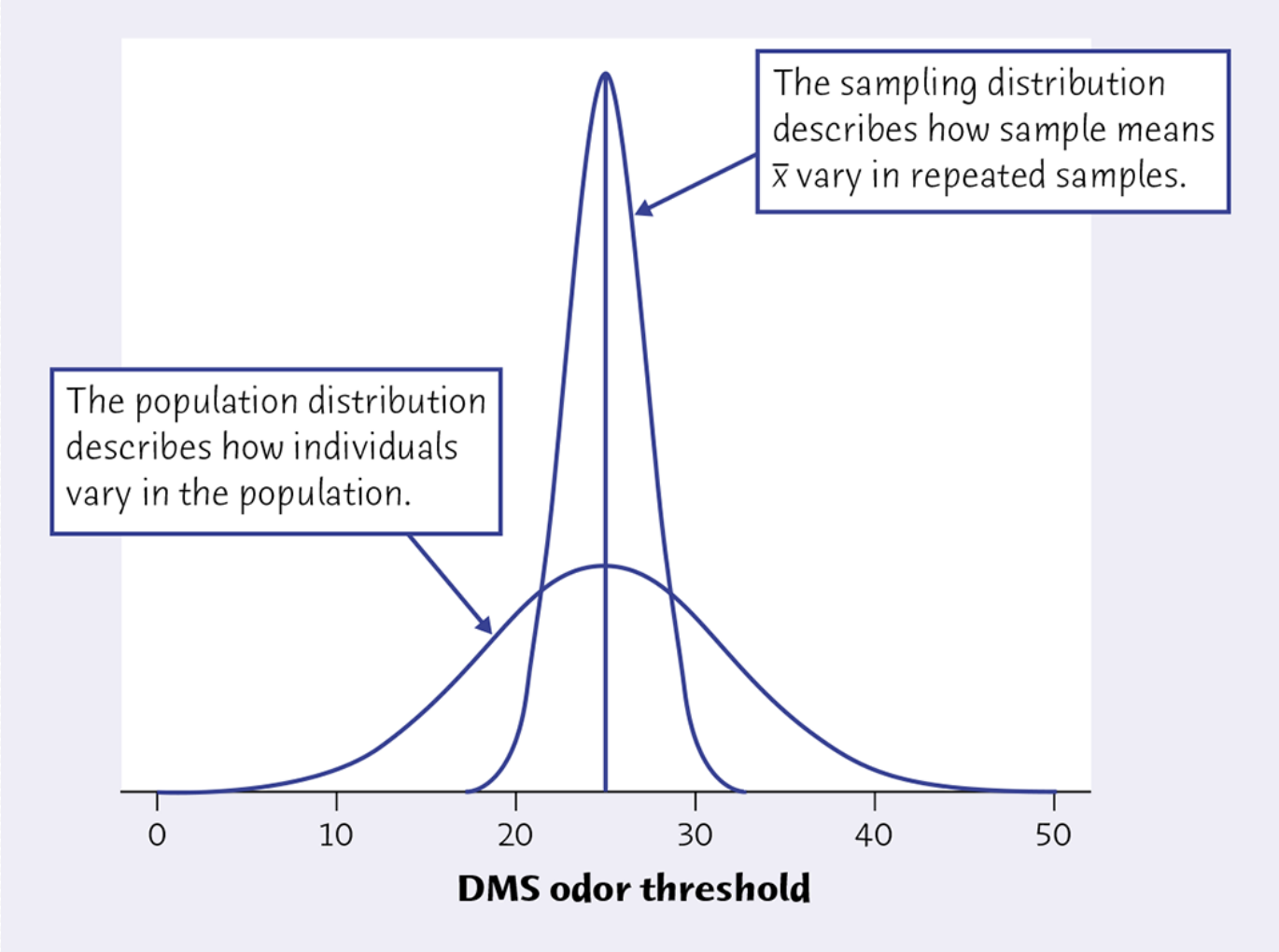
- When we use software to imitate chance behavior to carry out tasks such as exploring sampling distributions, this is called simulation.

Sampling distributions

- The population distribution of a variable is the distribution of values of the variable among all individuals in the population.
- The sampling distribution of a statistic is the distribution of values taken by the statistic in all possible samples of the same size from the same population.

Be careful: *The population distribution describes the individuals that make up the population. A sampling distribution describes how a statistic varies in many samples from the population.*

Population distributions vs. sampling distributions



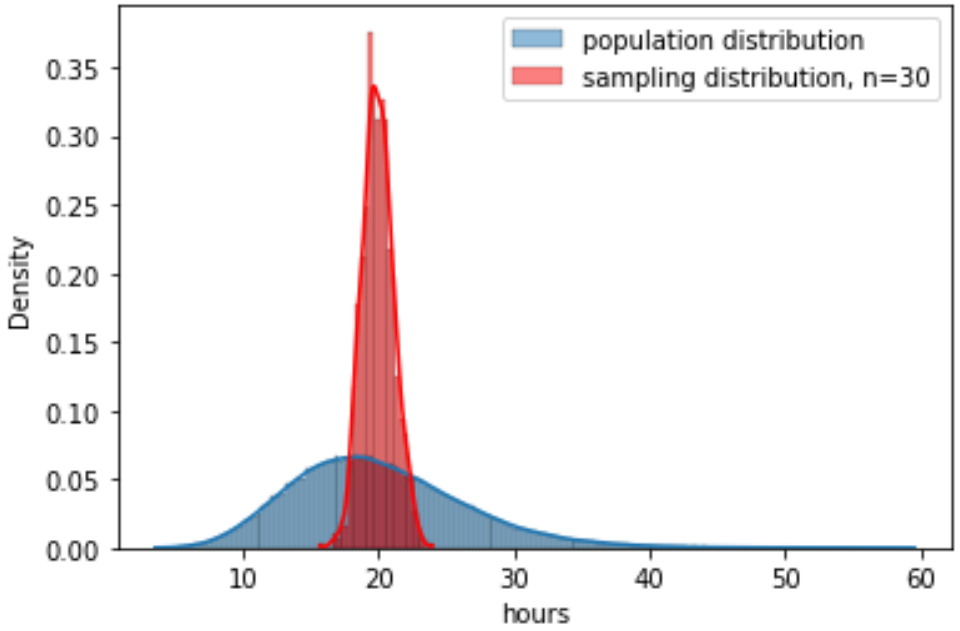
Example population vs. sampling distribution

```
# Function for creating a list of sample means
# The number of samples is "n_samples" and size of each sample is "sample_size".
# We calculate the mean of each sample and store all those sample mean values in the list "sample_means".
def sample_mean_calculator(population_array, sample_size, n_samples):
    sample_means = []
    for i in range(n_samples):
        sample = np.random.choice(population_array, size = sample_size, replace= False)
        sample_mean = np.mean(sample)
        sample_means.append(sample_mean)
    return sample_means
```

Example population vs. sampling distribution

```
# Calculating 1000 sample means from 1000 samples each of size 30
sample_means = sample_mean_calculator(df_hours['hours'], 30, 1000)
```

```
sns.histplot(df_hours['hours'], stat = 'density', kde = True,
             label = "population distribution")
sns.histplot(sample_means, stat = 'density', color = "red",
             label = "sampling distribution, n=30", kde = True)
plt.xlabel("hours")
plt.legend()
plt.show()
```



The sampling distribution of $\bar{x}$ (part I)

- When we choose many SRSs from a population, the sampling distribution of the sample mean is centered at the population mean μ and is less spread out than the population distribution.

MEAN AND STANDARD DEVIATION OF A SAMPLE MEAN

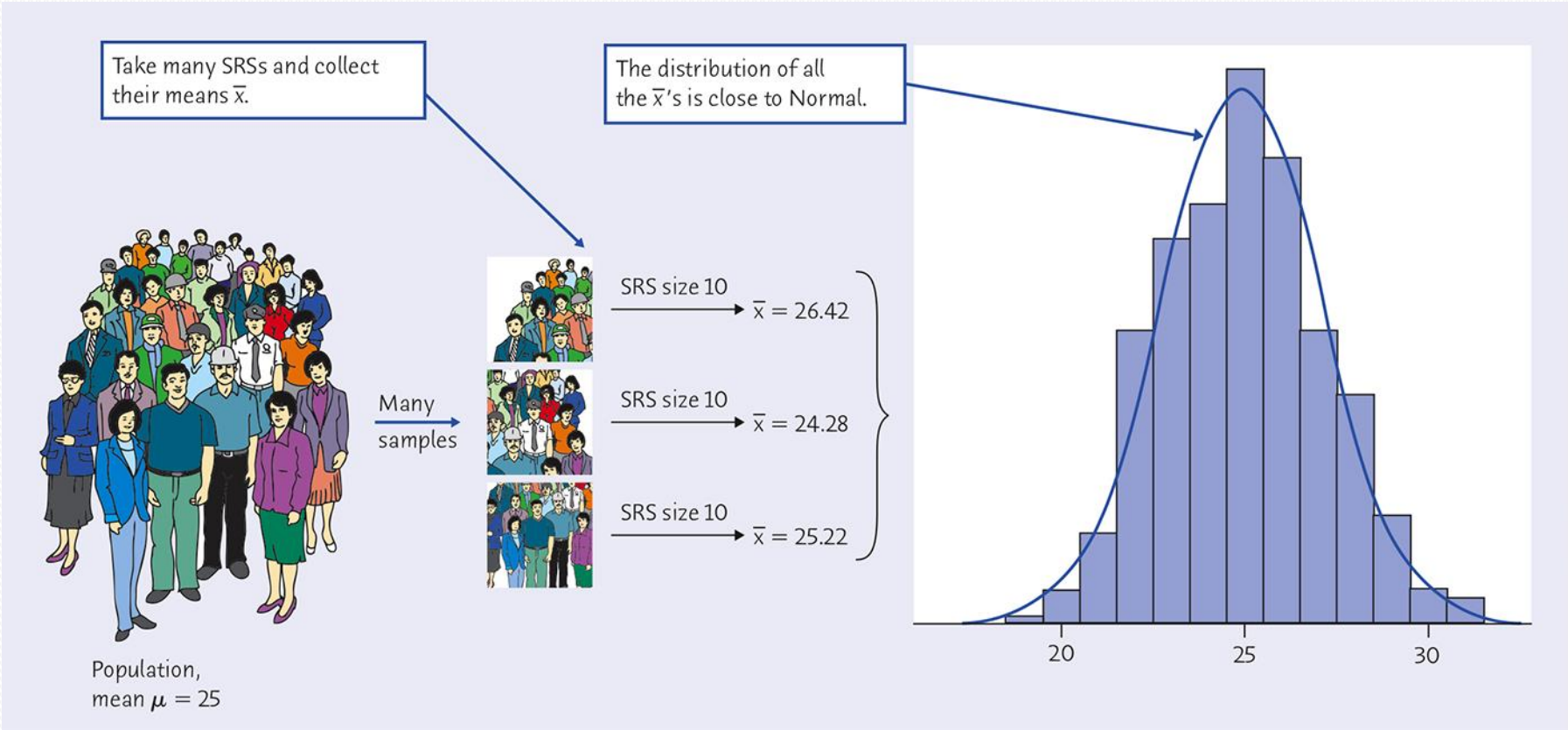
- Suppose that $\bar{x}$ is the mean of an SRS* of size n drawn from a large population with mean μ and standard deviation σ. Then the sampling distribution of $\bar{x}$ has **mean $\boldsymbol{\mu}$** and **standard deviation $\boldsymbol{\sigma/\sqrt{n}}$**.

- Because the mean of the statistic $\bar{x}$ is always equal to the mean μ of the population (that is, the sampling distribution of $\bar{x}$ is centered at μ), we say the statistic $\bar{x}$ is an **unbiased estimator** of the parameter μ.

  ***Note*:** On any particular sample, $\bar{x}$ may fall above or below μ.

* SRS = Simple Random Sample

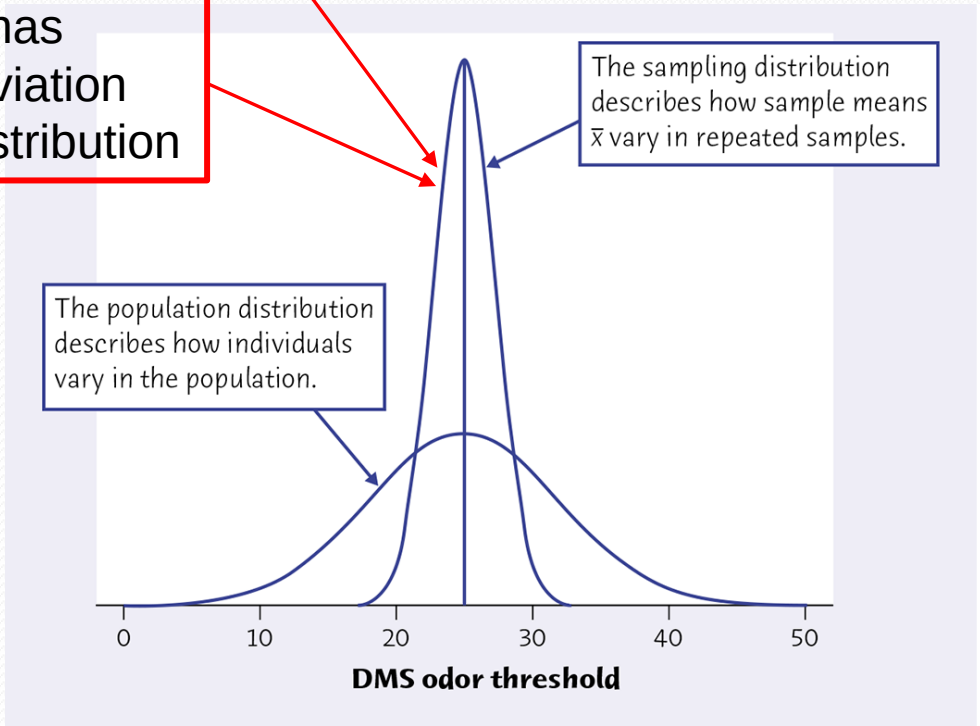
The sampling distribution of $\bar{x}$ (illustrated)



The sampling distribution of $\bar{x}$ (part II)

- Because the standard deviation of the sampling distribution of $\bar{x}$ is $\sigma/\sqrt{n}$, **the averages are less variable than individual observations**.

Sampling distribution has a smaller standard deviation than the population distribution



The sampling distribution of $\bar{x}$ (part II)

- Standard deviation of the sampling distribution of $\bar{x}$ is $\sigma/\sqrt{n}$.
- Not only is the standard deviation of the distribution of $\bar{x}$ smaller than the standard deviation of individual observations, **it gets smaller as we take larger samples**. The results of large samples are less variable than the results of small samples.

Note: *Although the standard deviation of the distribution of $\bar{x}$ gets smaller, it does so at the rate of $\sqrt{n}$, not n. To cut the sampling distribution's standard deviation in half, for instance, you must take a sample four times as large, not just twice as large.*

The sampling distribution of $\bar{x}$ (part III)

- We have described the center and variability of the sampling distribution of a sample mean $\bar{x}$, but not its shape. **The shape of the sampling distribution depends on the shape of the population distribution.**
- In one important case, there is a simple relationship between the two distributions: **if the population distribution is Normal, then so is the sampling distribution of the sample mean.**

The sampling distribution of $\bar{x}$ (part III)

SAMPLING DISTRIBUTION OF A SAMPLE MEAN

- If individual observations have the $N(\mu,\ \sigma)$ distribution, then the sample mean $\bar{x}$ of an SRS of size n has the $N(\mu,\ {}^{\sigma}/_{\sqrt{n}})$ distribution.

The central limit theorem

- Most population distributions are not Normal. What is the shape of the sampling distribution of sample means when the population distribution isn't Normal?

- A remarkable fact is that as the sample size increases, the distribution of sample means changes its shape: It looks less like that of the population and more like a Normal distribution!

The central limit theorem

- Draw an SRS of size n from any population with mean μ and finite standard deviation σ. The **central limit theorem** says that when *n* is large, the sampling distribution of the sample mean $\bar{x}$ is approximately Normal. That is,

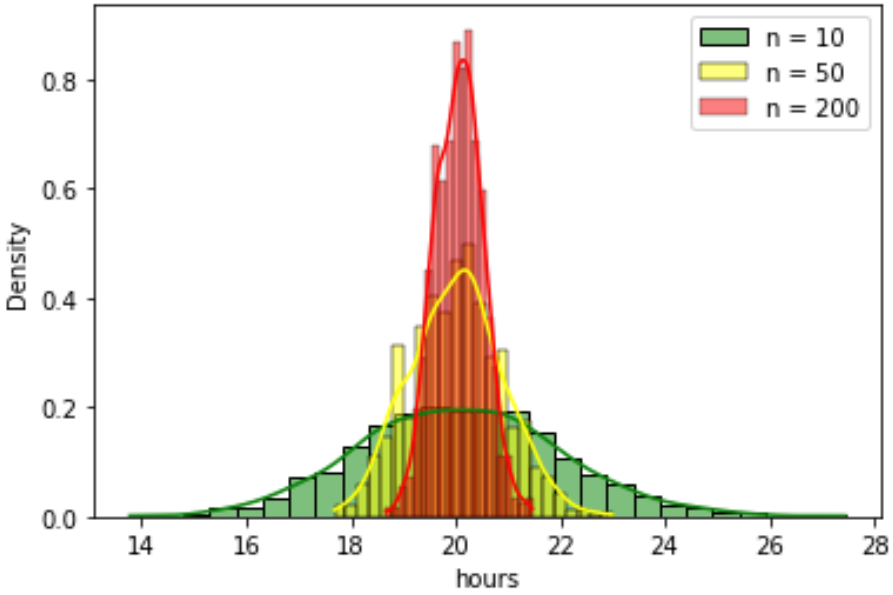
$$\bar{x} \text{ is approximately } N\left(\mu, {}^{\sigma}/_{\sqrt{n}}\right)$$

- The central limit theorem allows us to use Normal probability calculations to answer questions about sample means from many observations, even when the population distribution is not Normal.

The central limit theorem - example

```
sample_means_10 = sample_mean_calculator(df_hours['hours'], 10, 1000)
sample_means_50 = sample_mean_calculator(df_hours['hours'], 50, 1000)
sample_means_200 = sample_mean_calculator(df_hours['hours'], 200, 1000)
```

```
sns.histplot(sample_means_10, stat = 'density', color = "green", label = "n = 10", kde = True)
sns.histplot(sample_means_50, stat = 'density', color = "yellow", label = "n = 50", kde = True)
sns.histplot(sample_means_200, stat = 'density', color = "red", label = "n = 200", kde = True)
plt.xlabel("hours")
plt.legend()
plt.show()
```



Central limit theorem (example, 15.8 part I)

Based on service records from the past year, the time (in hours) that a technician requires to complete preventive maintenance on an air conditioner follows a distribution that is strongly right-skewed and whose most likely outcomes are close to 0 (see graph (a) in next slide). The mean time is $\mu = 1$ hour and the standard deviation is $\sigma = 1$.

Your company will service an SRS of 70 air conditioners. You have budgeted 1.1 hours per unit. What is the chance that the average hours exceed 1.1 hours?

The central limit theorem states that the sampling distribution of the mean time spent working on the 70 units has

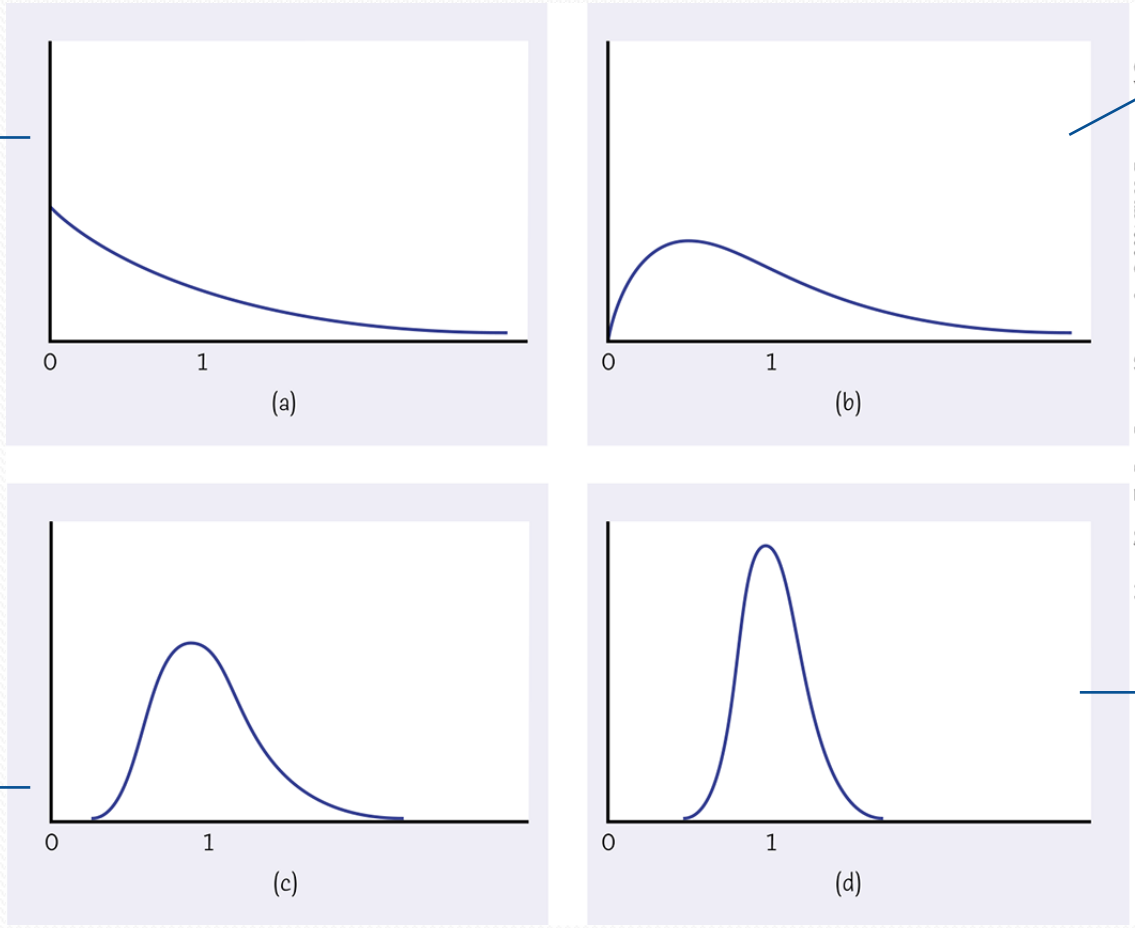
$$\mu_{\bar{x}} = \mu = 1,$$

$$\sigma_{\bar{x}} = \frac{\sigma_x}{\sqrt{n}} = \frac{1}{\sqrt{70}} = 0.12,$$

and a Normal distribution shape.

Central limit theorem in action (example 15.7)

Population distribution of time to repair an air conditioner, an exponential distribution with mean 1 and std 1



Mean sampling distribution for samples of size n=2

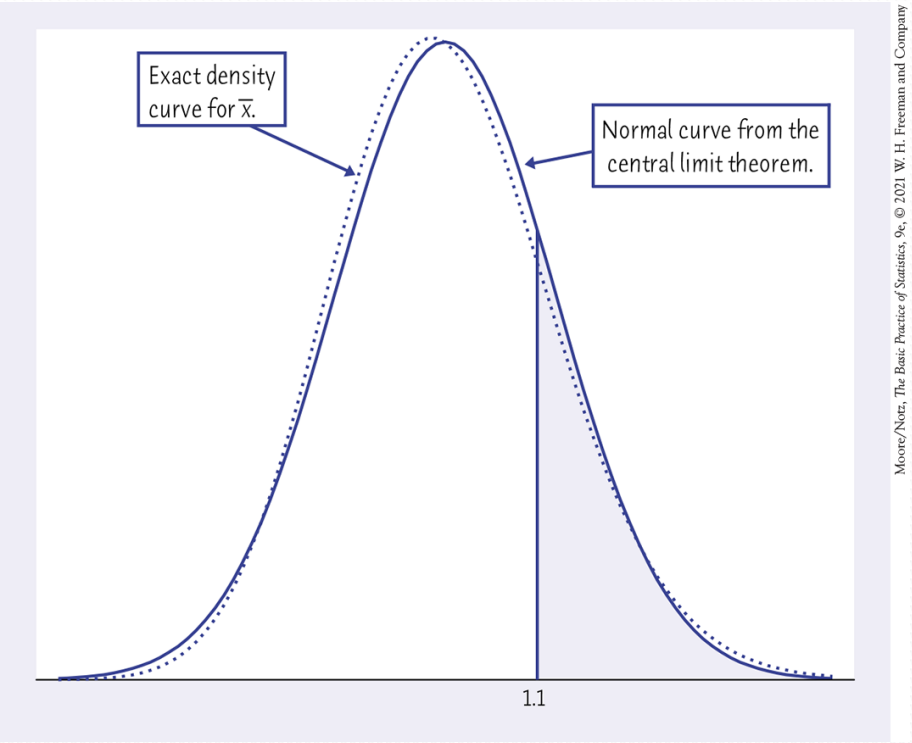
Mean sampling distribution for samples of size n=10

Mean sampling distribution for samples of size n=25

Central limit theorem (example, part II)

Your company will service an SRS of 70 air conditioners. You have budgeted 1.1 hours per unit. What is the chance that the average hours exceed 1.1 hours?

The sampling distribution of the mean time spent working is approximately $N(1, 0.12)$ because $n = 70 \geq 30$.



Moore/Notz, *The Basic Practice of Statistics*, 9e, © 2021 W. H. Freeman and Company

$$z = \frac{1.1 - 1}{0.12} = 0.83$$

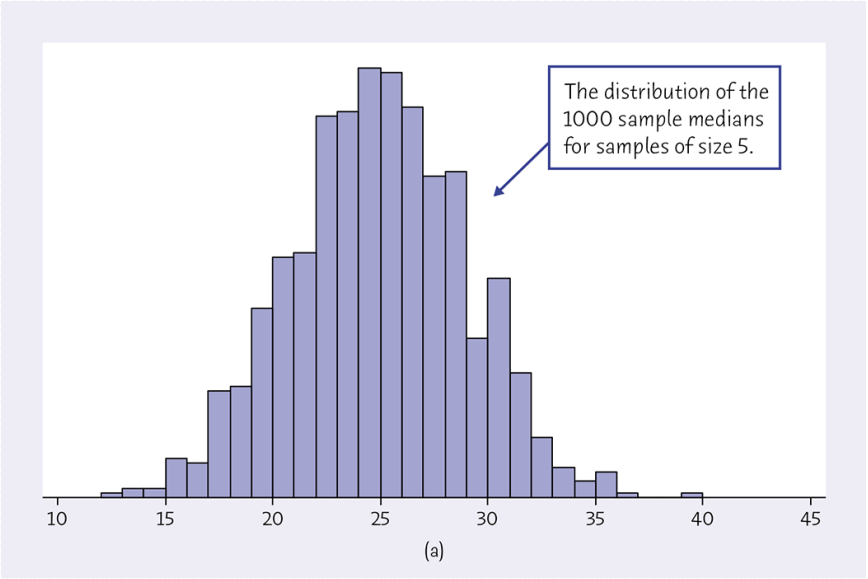
$$P(\bar{x} > 1.1) = P(Z > 0.83)$$

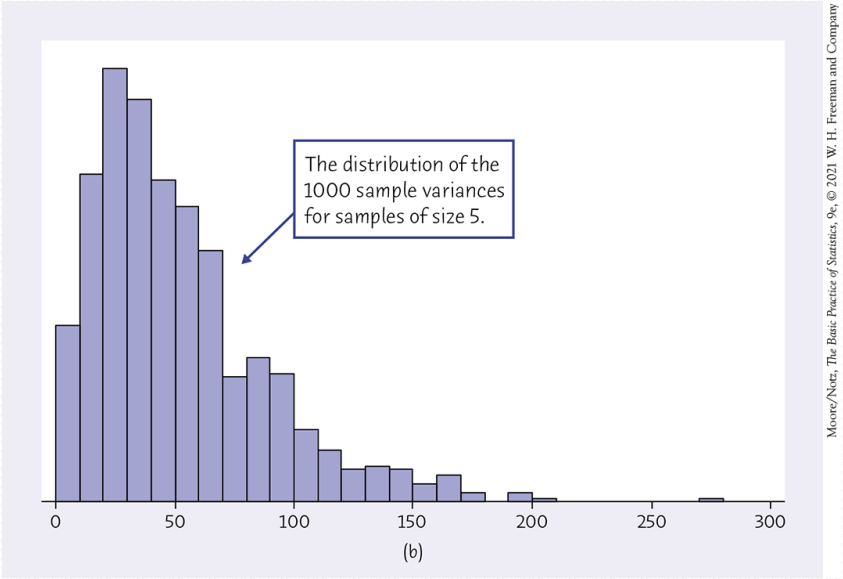
$$= 1 - 0.7967 = 0.2033$$

If you budget 1.1 hours per unit, there is a 20% chance that the technicians will not complete the work within the budgeted time.

Sampling distributions and statistical significance (part I)

- We have looked carefully at the sampling distribution of a sample mean.
- However, any statistic we can calculate from a sample will have a sampling distribution.





Sampling distributions and statistical significance (part II)

- The sampling distribution of a sample statistic is determined by the particular sample statistic we are interested in, the distribution of the population of individual values from which the sample statistic is computed, and the method by which samples are selected from the population.
- The sampling distribution allows us to determine the probability of observing any particular value of the sample statistic in another such sample from the population. An observed effect so large that it would rarely occur by chance is called **statistically significant**.
- Consider the second graph on the previous slide. We may decide, based on our observed set of 1000 samples, that because we see only 2 with variances above 200, this is a statistically significant event.

Example – statistical significance

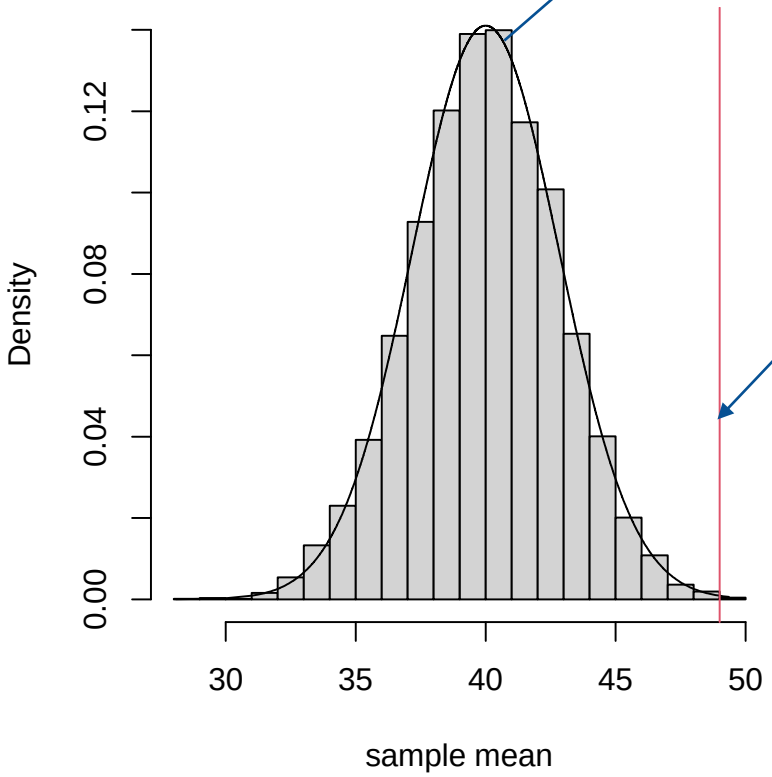
Suppose there is a hypothesis that the true mean of a population is 40. Suppose you know that the population standard deviation is 20.

You have data from a SRS of size 50 from this population and $\bar{x}$ is 49. Do you have enough statistical evidence to reject the null hypothesis that the true population mean is 40?

Example – statistical significance

Mean sampling distribution under the hypothesis that the true population mean is 40 → Normal(40, 20/sqrt(50))

We know this, not part of the hypothesis

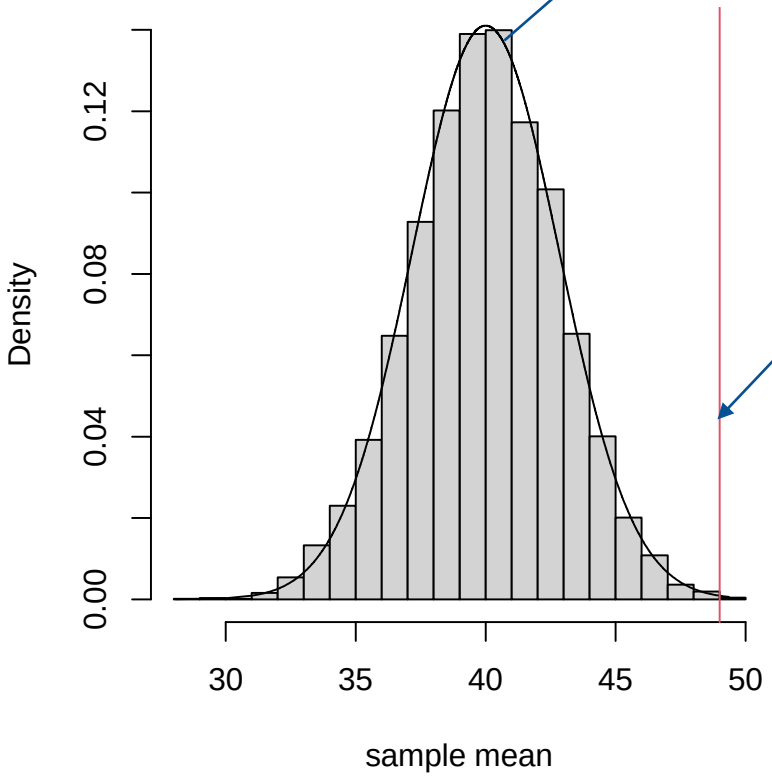


An observed sample mean of 49 seems to be unusual if the hypothesis was true.

Example – statistical significance

Mean sampling distribution under the hypothesis that the true population mean is 40 → Normal(40, 20/sqrt(50))

We know this, not part of the hypothesis



An observed sample mean of 49 seems to be unusual if the hypothesis was true.

What is $P(\bar{x} > 49$ under the hypothesis that $\mu = 40)$?

Z =(49 - 40)/20/sqrt(50) = 3.181981

$P(Z > 3.181981) = 0.0007$

→ We do have enough evidence to reject the hypothesis that $\mu = 40$. In other words, the observed difference in means is **statistically significant**.