Chapter 8

Fundamental Sampling Distributions and Data Descriptions

Section 8.1

Random
Sampling

Definition 8.1

A **population** consists of the totality of the observations with which we are concerned.

Definition 8.1

A **population** consists of the totality of the observations with which we are concerned.

- We want to draw useful conclusions from observations.
- What is the most popular coffee brand? What is the most common disease for > 70yo? What is the mean height of elephants in India?

Definition 8.1

A **population** consists of the totality of the observations with which we are concerned.

- Problem: we cannot make measurements over all the population
  - time, cost, etc.
- Instead, we take a subset of population, and draw conclusion from the subset
  - called inference

Definition 8.2

A **sample** is a subset of a population.

Definition 8.2

A **sample** is a subset of a population.

- In order for inference to be valid, a sample should represent the population well
- What if I take samples from close friends for the distribution of occupation?
  - Not likely to represent the whole population
  - overestimation/underestimation possible-called bias

Definition 8.3

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

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

Section 8.2

Some Important
Statistics

Definition 8.4

Any function of the random variables constituting a random sample is called a **statistic**.

Definition 8.4

Any function of the random variables constituting a random sample is called a **statistic**.

- n random samples: n random variables
- statistic: function of random samples, so it is a random variable

Definition 8.4

> Any function of the random variables constituting a random sample is called a **statistic**.

- We want to estimate some properties of population from statistic,
  - what is the mean of a property of population?
  - what is its variance?

Definition: sample mean

Let $X_1, X_2, \ldots, X_n$ represent n random variables.

Sample mean:

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i\,.$$

Definition: sample variance

Let $X_1, X_2, \ldots, X_n$ represent n random variables.

Sample variance:

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2.$$

example 8.2

Example 8.2: A comparison of coffee prices at 4 randomly selected grocery stores in San Diego showed increases from the previous month of 12, 15, 17, and 20 cents for a 1-pound bag. Find the variance of this random sample of price increases.

Theorem 8.1

If S^2 is the variance of a random sample of size n, we may write

$$S^2 = \frac{1}{n(n-1)}\left[n\sum_{i=1}^{n} X_i^2 - \left(\sum_{i=1}^{n} X_i\right)^2\right].$$

Section 8.4

Sampling Distribution of Means and the Central Limit Theorem

Definition 8.5

The probability distribution of a statistic is called a **sampling distribution**.

- Statistic: function of random samples,
so it is a random variable, and has distribution.

- Interested in the distribution of sample mean, sample variance, etc.

Definition 8.5

The probability distribution of a statistic is called a **sampling distribution**.

Let $X_1, X_2, \ldots, X_n$ represent ***n* random samples**

Sample mean:

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

What is the distribution of sample mean?

Of course, it depends on distribution of X_i

But for large n, sample mean follows a particular distribution!

sampling distribution of mean

Example: when sample itself is normally distributed

Each observation X_i, $i = 1, 2, \ldots, n$, **has** **normal distribution**

$$\bar{X} = \frac{1}{n}(X_1 + X_2 + \cdots + X_n)$$

has a normal distribution with mean

$$\mu_{\bar{X}} = \frac{1}{n}(\underbrace{\mu + \mu + \cdots + \mu}_{n \text{ terms}}) = \mu \text{ and variance } \sigma^2_{\bar{X}} = \frac{1}{n^2}(\underbrace{\sigma^2 + \sigma^2 + \cdots + \sigma^2}_{n \text{ terms}}) = \frac{\sigma^2}{n}.$$

What if X_i are not normal?

Sample mean still has normal dist. for large n!

Theorem 8.2

Central Limit Theorem: If $\bar{X}$ is the mean of a random sample of size n taken from a population with mean μ and finite variance σ^2, then the limiting form of the distribution of

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}},$$

as $n \to \infty$, is the standard normal distribution $n(z; 0, 1)$.

Theorem 8.2

$$Z = \sum_{i=1}^{n} \frac{1}{\sqrt{n}\sigma} Y_i, \quad Y_i = X_i - \mu$$

using Taylor's formula,

$$f(t) = f(0) + f'(0)t + f''(0)\frac{t^2}{2!} + \cdots = \sum_{k=0}^{\infty} f^{(k)}(0)\frac{t^k}{k!}$$

Theorem 8.2

$$Z = \sum_{i=1}^{n} \frac{1}{\sqrt{n}\sigma} Y_i, \quad Y_i = X_i - \mu$$

using Taylor's formula,

$$\begin{aligned} M_{Y_i/\sqrt{n}\sigma}(t) &= M_{Y_i}(t/\sqrt{n}\sigma) \\ &= M_{Y_i}(0) + M'_{Y_i}(0)\frac{t}{\sqrt{n}\sigma} + M''_{Y_i}(0)\frac{1}{2!}\left(\frac{t}{\sqrt{n}\sigma}\right)^2 + M'''_{Y_i}(0)\frac{1}{3!}\left(\frac{t}{\sqrt{n}\sigma}\right)^3 \dots \\ &= 1 + \frac{\sigma^2 t^2}{2n\sigma^2} + O(\frac{1}{n\sqrt{n}}) \end{aligned}$$

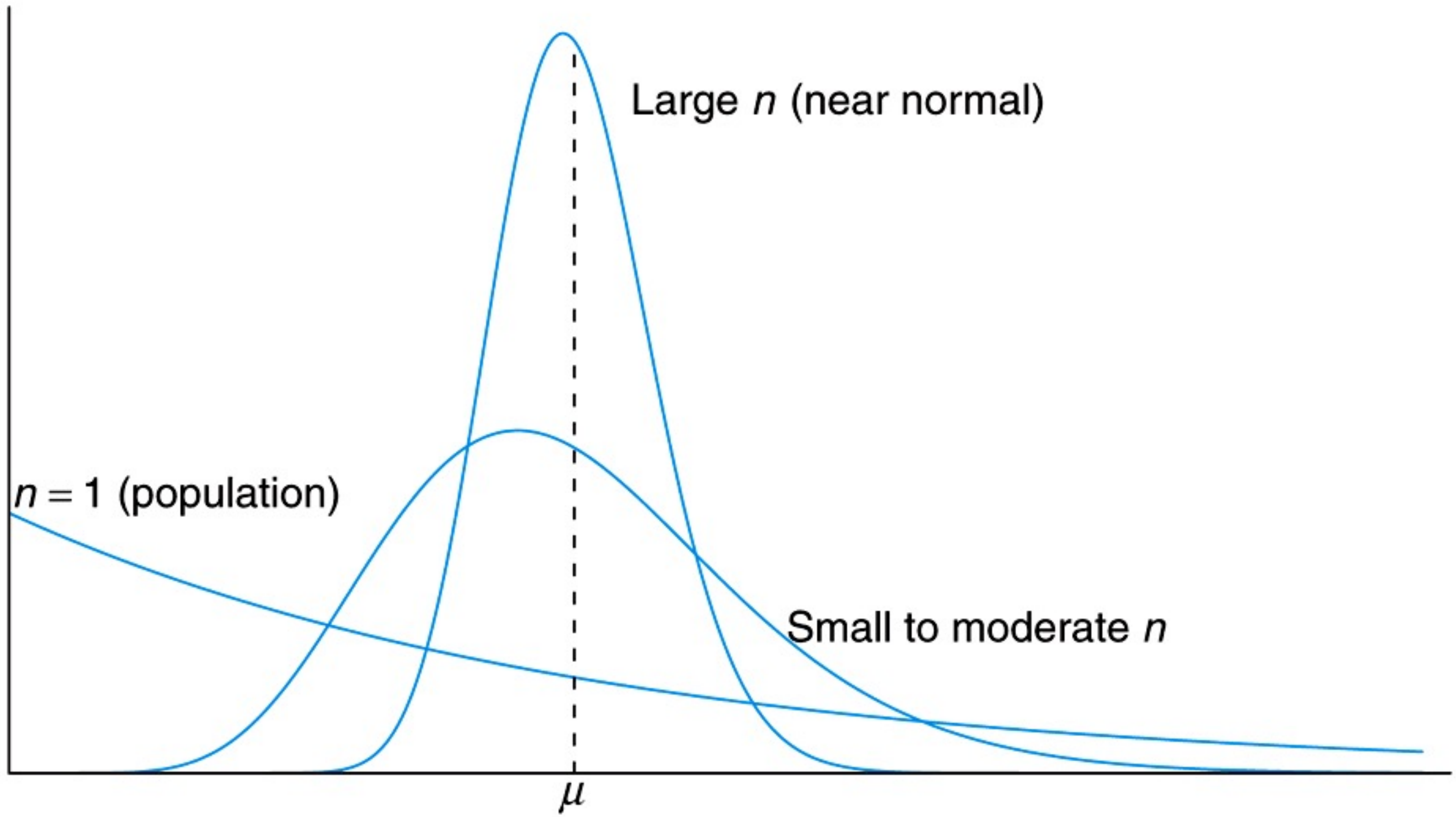
Theorem 8.2

$$M_Z(t) = M_{Y_i/\sqrt{n}\sigma}(t)^n = \left(1 + \frac{t^2}{2n} + O(n^{-\frac{3}{2}})\right)^n \to \exp\left(\frac{t^2}{2}\right) \text{ as } n \to \infty$$

This is MGF of N(0,1)!

By uniqueness of MGFs, the distribution must be standard Gaussian

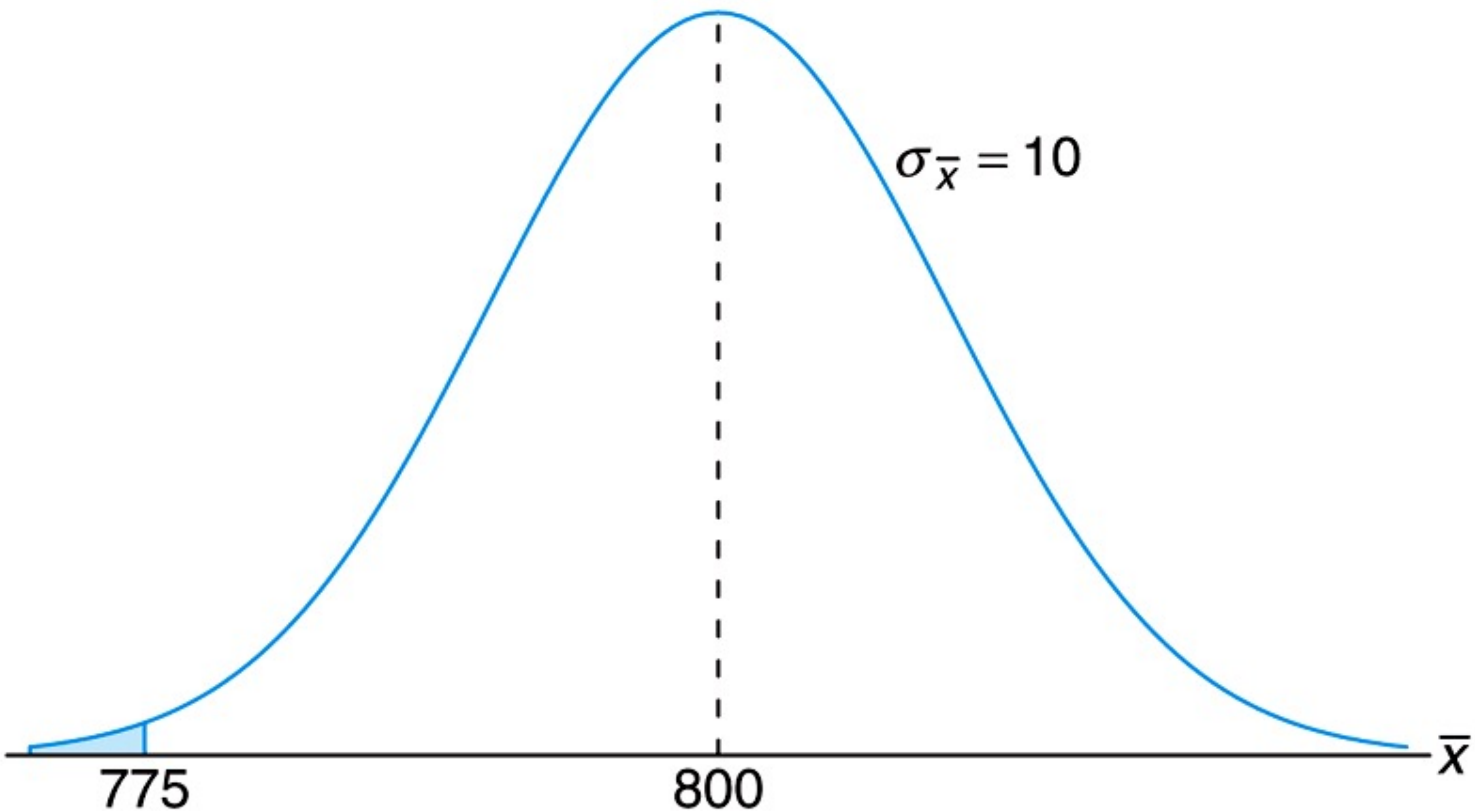
Figure 8.1 Illustration of the Central Limit Theorem (distribution of $\bar{X}$ for $n = 1$, moderate n, and large **n**)



example 8.4

Example 8.4: An electrical firm manufactures light bulbs that have a length of life that is approximately normally distributed, with mean equal to 800 hours and a standard deviation of 40 hours. Find the probability that a random sample of 16 bulbs will have an average life of less than 775 hours.

Figure 8.2 Area for Example 8.4



Comments on CLT

- Very important and useful
  - Oftentimes we are mostly interested in sample mean
  - If we get enough samples, we know the good approximation distribution of the sample mean!
  - Explains why Gaussian is ‘normal’ distribution: if a phenomenon (error) occurs in ‘additive’ way, they look approximately Gaussian

Comments on CLT

- CLT says **sample mean** becomes Gaussian, but NOT sample themselves
  - Samples has distribution from population
  - CLT applies to sample means (each of size n) measured many times
  - How large should n be?
  - Typically 20~30

Section 8.5

Sampling
Distribution of S^2

Definition: sample variance

The probability distribution of a statistic is called a **sampling distribution**.

Let $X_1, X_2, \ldots, X_n$ represent n random variables.

Sample variance:

$$S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2.$$

What is the distribution of sample variance?

Mostly interested in case for finite n and normally distributed X_i

Good approximation when X_i is approximately normal

Theorem 8.4

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

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

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

assume variance is known but mean is unknown

Derivation

$$\begin{aligned}
\sum_{i=1}^{n}(X_i-\mu)^2 &= \sum_{i=1}^{n}(X_i-\bar{X}+\bar{X}-\mu)^2 \\
&= \sum_{i=1}^{n}(X_i-\bar{X})^2+\sum_{i=1}^{n}(\bar{X}-\mu)^2+2(\bar{X}-\mu)\sum_{i=1}^{n}(X_i-\bar{X}) \\
&= (n-1)S^2+n(\bar{X}-\mu)^2
\end{aligned}$$

Next we show S^2 and $\bar{X}$ are independent

Derivation

1. $\bar{X}$ and $X_i - \bar{X}$ are jointly normal (means, both can be written as a linear combinations of independent Gaussian RVs)
2. $\bar{X}$ and $X_i - \bar{X}$ are uncorrelated
3. $\bar{X}$ and $X_i - \bar{X}$ are independent (because they are jointly normal & uncorrelated!)

Derivation

$$\sum_{i=1}^{n}\left(\frac{X_i-\mu}{\sigma}\right)^2 = \frac{(n-1)S^2}{\sigma^2} + \frac{(\bar{X}-\mu)^2}{\sigma^2/n}$$

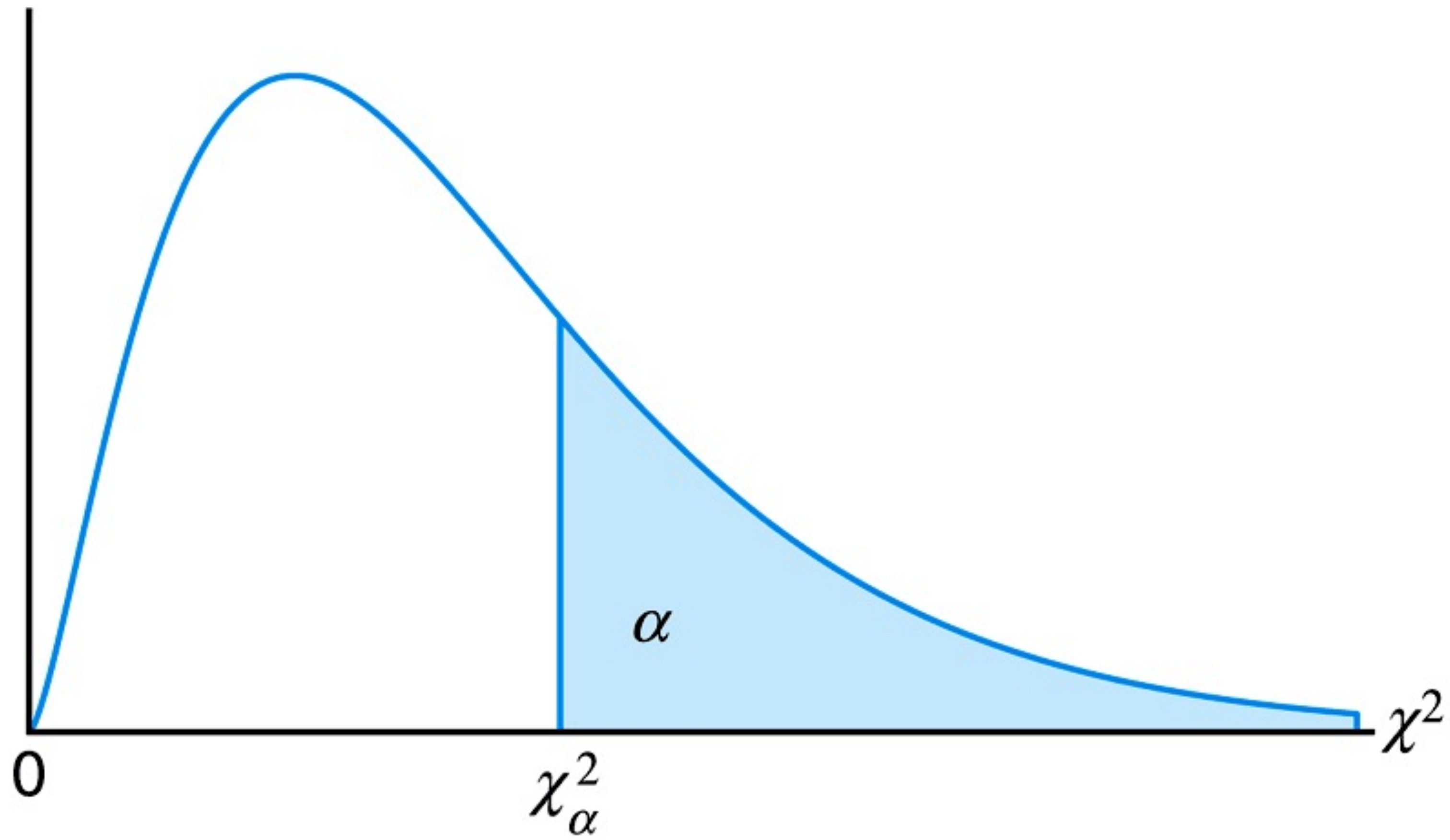
taking MGF at both sides,

$$(1-2t)^{-n} = M_{\frac{(n-1)S^2}{\sigma^2}}(t)(1-2t)^{-1}$$

This means

$\frac{(n-1)S^2}{\sigma^2}$ has chi-square distribution with $n-1$ degree of freedom!

Figure 8.7 The chi-squared distribution



Section 8.6

t-Distribution

Theorem 8.5

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

$$T = \frac{Z}{\sqrt{V/v}},$$

is given by the density function

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

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

Corollary 8.1

Let $X_1, X_2, \ldots, X_n$ be independent random variables that are all normal with mean μ and standard deviation σ. Let

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \qquad \text{and} \qquad S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2.$$

Then the random variable $T = \frac{\bar{X}-\mu}{S/\sqrt{n}}$ has a t-distribution with $v = n - 1$ degrees of freedom.

General statement

Let $X_1, X_2, \ldots, X_n$ i.i.d random variables with mean μ and finite standard deviation. Let

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \qquad S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

Then $T = \frac{\bar{X}-\mu}{S/\sqrt{n}}$ becomes $N(0,1)$ as $n \to \infty$.

Shows CLT holds with sample variance
t-distribution becomes Gaussian for large n

Figure 8.8 The *t*-distribution curves for v = 2, 5, and ∞

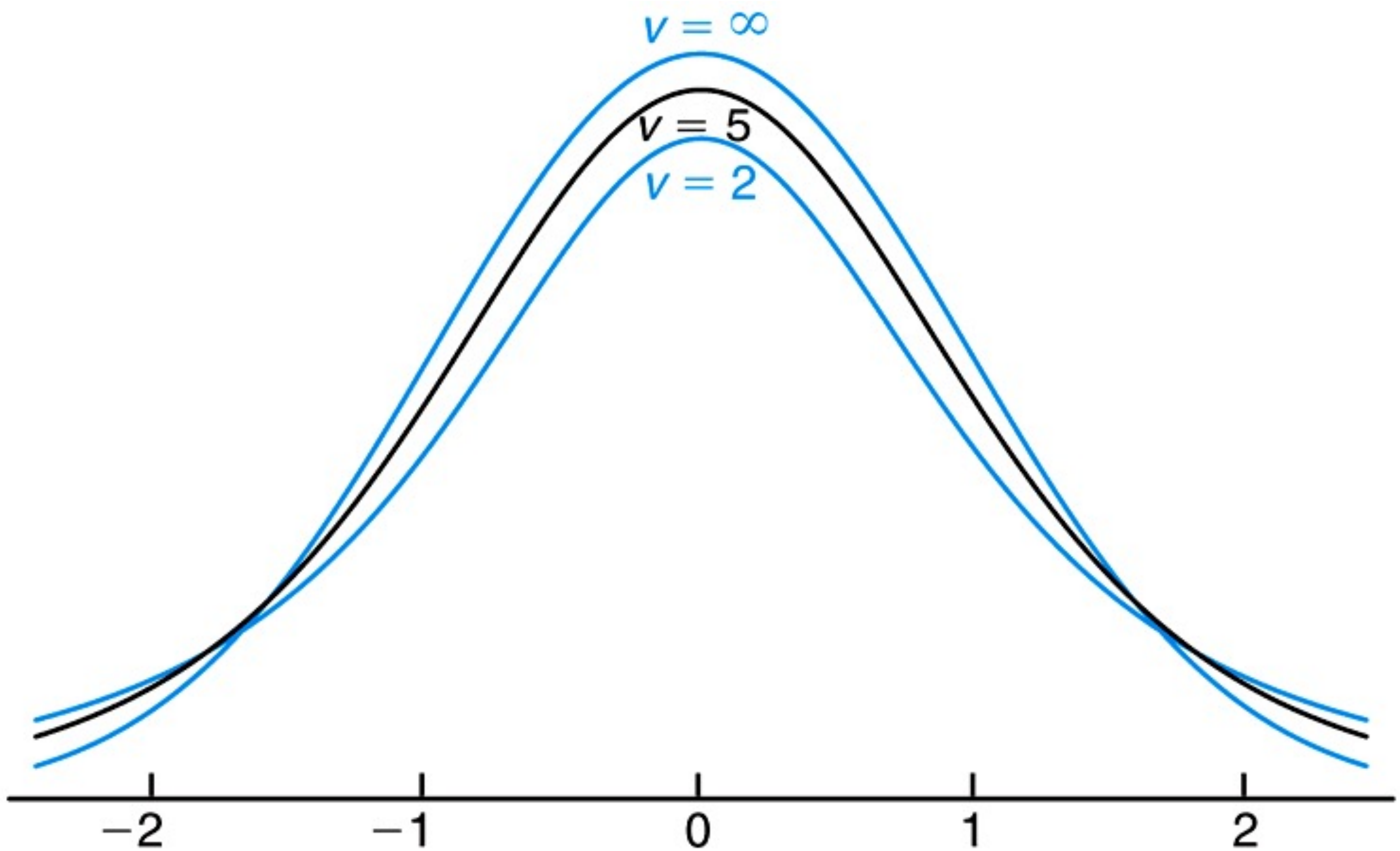
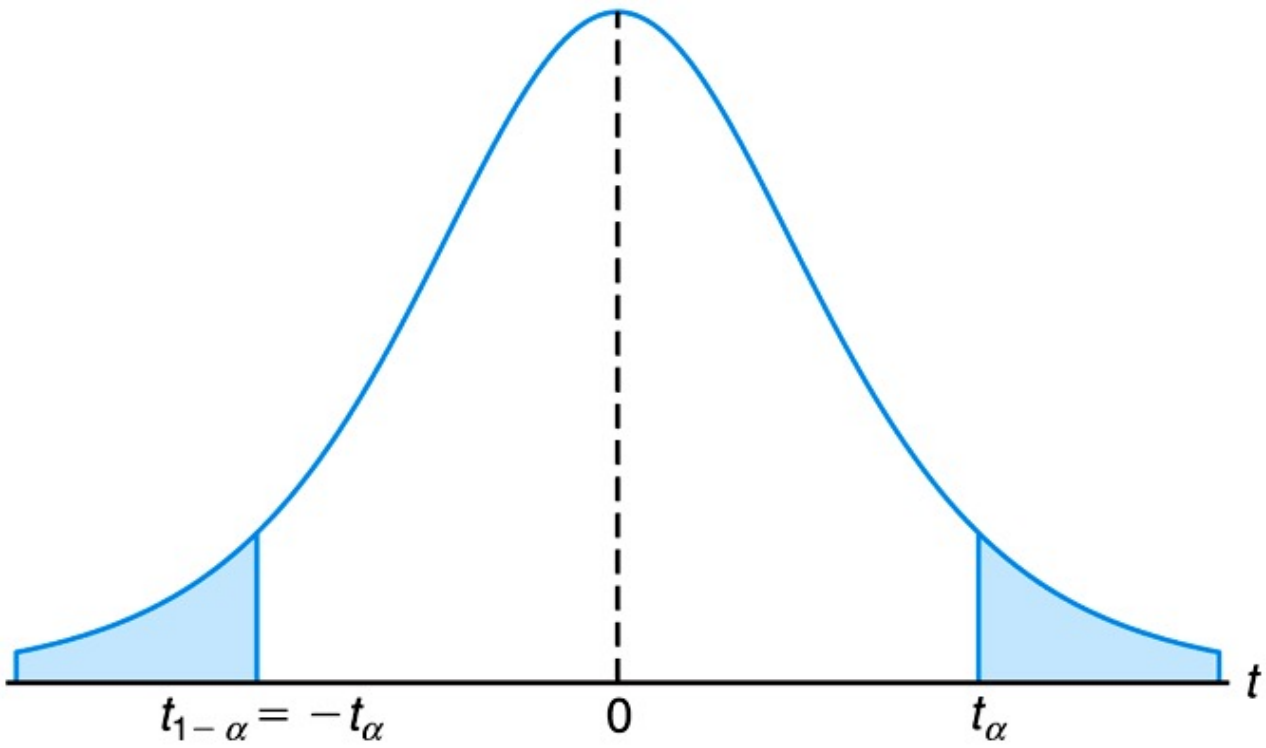


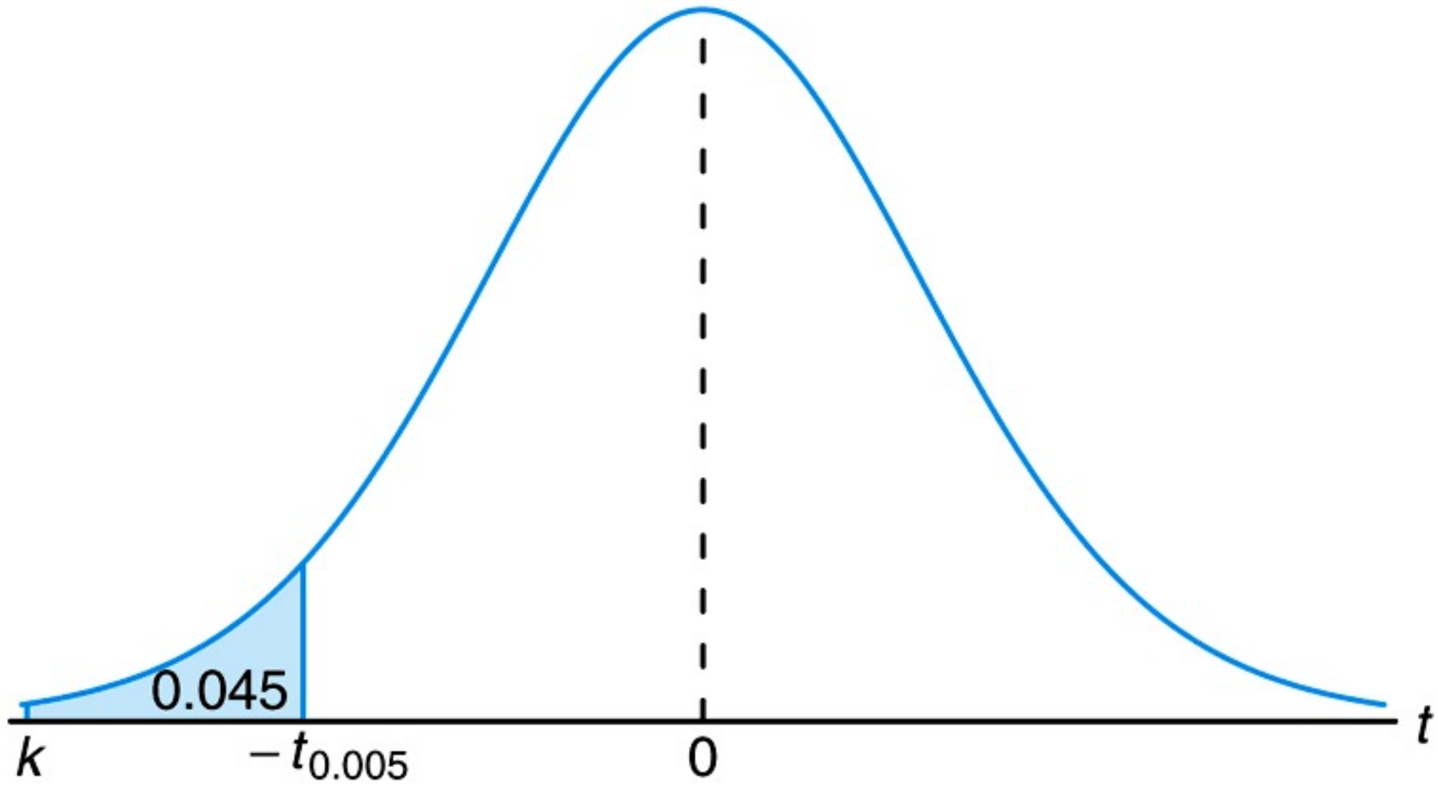
Figure 8.9 Symmetry property (about 0) of the *t*-distribution



Example 8.10

Example 8.10: Find k such that $P(k < T < -1.761) = 0.045$ for a random sample of size 15 selected from a normal distribution and $\frac{\overline{X}-\mu}{s/\sqrt{n}}$.

Figure 8.10 The *t*-values for Example 8.10



Section 8.7

F-Distribution

Motivation

- Interested in comparing two populations
  - Pharmacy company tested drug A and B to different subject groups. How different are the effects?
- Comparing sample statistics
  - Sample mean, sample variance
- F-distribution: comparing two sample variances
  - Which of two groups has larger variance?

F-distribution

Let U and V be two independent chi-square distributed RVs with v_1 and v_2 degrees of freedom. Then $F = \frac{U/v_1}{V/v_2}$ has F-distribution with v_1 and v_2 degrees of freedom with density

$$h(f) = \begin{cases} \frac{\Gamma[(v_1+v_2)/2](v_1/v_2)^{v_1/2}}{\Gamma(v_1/2)\Gamma(v_2/2)} \frac{f^{(v_1/2)-1}}{(1+v_1 f/v_2)^{(v_1+v_2)/2}}, & f > 0 \\ 0, & f \le 0 \end{cases}$$

Can be derived using function of RVs
Don’t need to memorize PDF..

F-distribution

f_α: value of f such that area under $h(f)$ to the right of f is equal to α

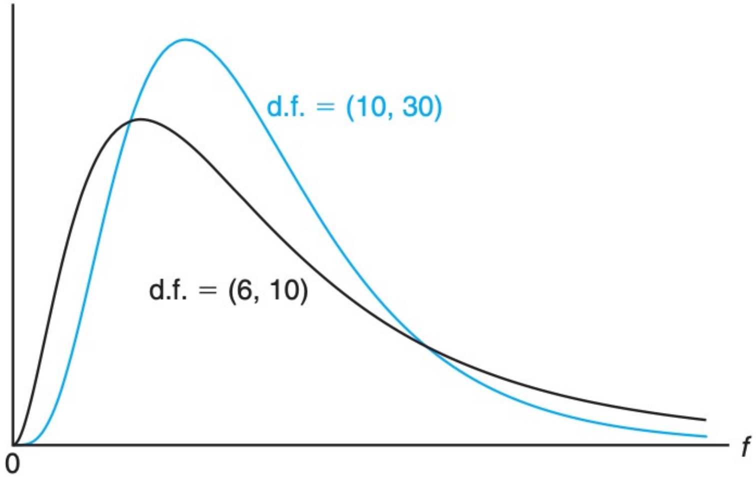


Figure 8.11: Typical F-distributions.

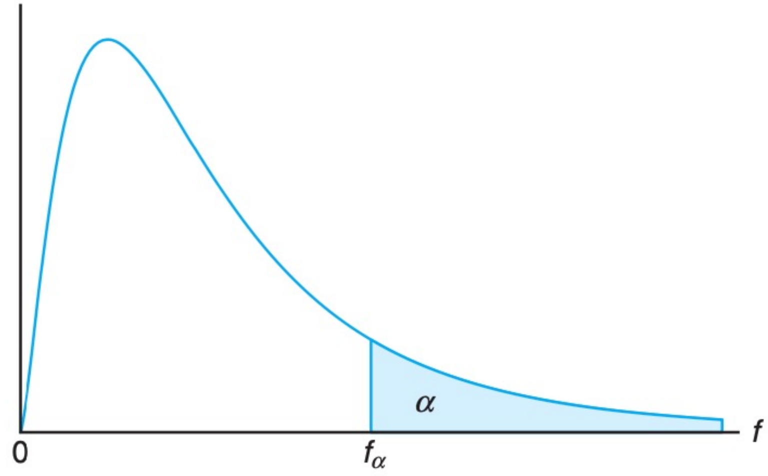


Figure 8.12: Illustration of the f_α for the F-distribution.

F-distribution

Theorem 8.8: If S_1^2 and S_2^2 are the variances of independent random samples of size n_1 and n_2 taken from normal populations with variances σ_1^2 and σ_2^2, respectively, then

$$F = \frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2} = \frac{\sigma_2^2 S_1^2}{\sigma_1^2 S_2^2}$$

has an F-distribution with $v_1 = n_1 - 1$ and $v_2 = n_2 - 1$ degrees of freedom.

From the definitions of sample variance and chi-square dist.