STA255: Statistical Theory

Chapter 7: Sampling Distributions and the Central Limit Theorem

Summer 2017

Introduction

- A population consists of the entire collection of the observations with which we are concerned.
- A sample is a subset of a population.
- A numerical summary of a population is called a parameter. In practice it is unknown.
- A numerical summary of a sample is called a statistic. For example, the mean, the sample variance, etc.
- The statistic varies from sample to sample and hence it is a random variable and has a probability distribution called the sampling distribution.
- The knowledge of the sampling distribution of a statistic helps to make inference about the corresponding population (true) parameter.

Sampling Distributions

- The random variables $Y_1, Y_2, \ldots, Y_n$ are said to form a random sample of size n if
  - The Y_i's are independent random variables.
  - Every Y_i has the same probability distribution.
- For a random sample, we write i.i.d.

$$Y_1, \ldots, Y_n \overset{i.i.d.}{\sim} f(y) \quad \text{(continuous)}$$

$$Y_1, \ldots, Y_n \overset{i.i.d.}{\sim} p(y) \quad \text{(discrete)}$$

- i.i.d: independent and identically distributed.

Example

The goal is to study the sampling distribution of:

- **Sample mean:** $\overline{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$.
- **Sample variance:** $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \overline{Y})^2$.

Sampling Distribution of $\overline{Y}$

Theorem

If $Y_1, Y_2, \ldots, Y_n \overset{i.i.d}{\sim} N(\mu, \sigma^2)$, *then*

$$\overline{Y} \sim N(\mu, \sigma^2/n)$$

Note:

$$Z = \frac{\overline{Y} - \mu}{\sigma/\sqrt{n}} \sim N(0,1)$$

Example: #7.11

A forester studying the effects of fertilization on certain pine forests in the Southeast is interested in estimating the average basal area of pine trees. In studying basal areas of similar trees for many years, he has discovered that these measurements (in square inches) are normally distributed with standard deviation approximately 4 square inches. If the forester samples $n = 9$ trees, find the probability that the sample mean will be within 2 square inches of the population mean.

χ^2 distribution

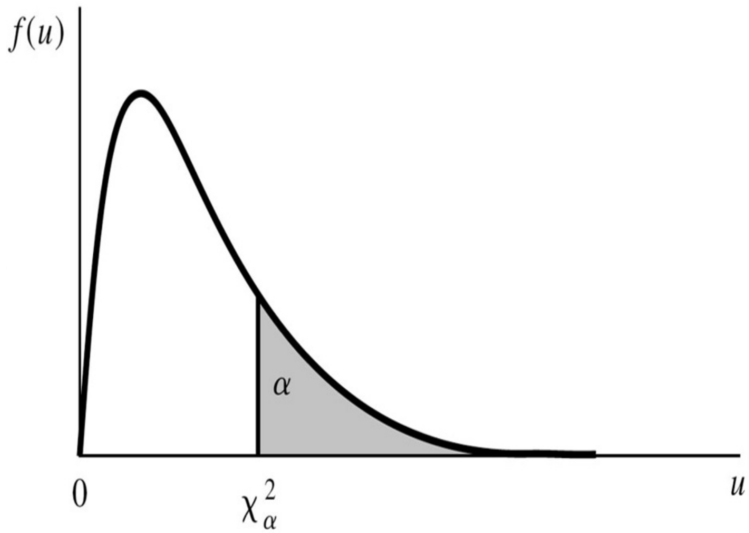
Theorem

Let $Y_1, \ldots, Y_n \overset{i.i.d.}{\sim} N(\mu, \sigma)$. *If* $Z_i = \frac{Y_i - \mu}{\sigma}$, *then*

$$\sum_{i=1}^{n} Z_i^2$$

has a χ^2 *distribution with n degrees of freedom (df).*

- We write χ_n or $\chi(n)$.
- Values of χ^2_n are given in Table 6 (p. 850)



Example

If $Z_1, Z_2, \ldots, Z_6$ denotes a random sample from the standard normal distribution, find a number b such that

$$P(\sum_{i=1}^{6} Z_i^2 < b) = 0.95$$

Sampling Distribution of S^2

Theorem 7.3

Let $Y_1, Y_2, \ldots, Y_n$ be a random sample from a normal distribution with mean μ and variance σ^2. Then

$$\frac{(n-1)S^2}{\sigma^2} = \sum_{i=1}^{n} \frac{(Y_i - \bar{Y})^2}{\sigma^2} \sim \chi^2_{n-1}$$

Also $\bar{Y}$ and S^2 are independent random variables.

Example: # 7.37

Let $Y_1, Y_2, \ldots, Y_5$ be a random sample of size 5 from a normal population with mean 0 and variance 1 and let $\bar{Y} = \frac{1}{5}\sum_{i=1}^{5} Y_5$. Let Y_6 be another independent observation from the same population. What is the distribution of

(a) $W = \sum_{i=1}^{5} Y_i^2$. Why?

(b) $U = \sum_{i=1}^{5}(Y_i - \bar{Y})^2$. Why?

(c) $\sum_{i=1}^{5}(Y_i - \bar{Y})^2 + Y_6^2$. Why?

t-Distribution

- So far, it was assumed that the population standard deviation σ is known.
- This assumption may be unreasonable.
- We want estimate both μ and σ.
- A natural statistic to consider to deal with inferences on μ is

$$T = \frac{\bar{Y} - \mu}{S/\sqrt{n}}$$

t-Distribution

- Note that

$$T = \frac{\bar{Y} - \mu}{S/\sqrt{n}} = \frac{(\bar{Y} - \mu)/(\sigma/\sqrt{n})}{\sqrt{S^2/\sigma^2}} = \frac{Z}{\sqrt{W/(n-1)}},$$

where

$$Z = \frac{\bar{Y} - \mu}{\sigma/\sqrt{n}} \sim N(0,1),$$

and

$$W = \frac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1).$$

- Thus

$$T = \frac{N(0,1)}{\sqrt{\chi^2_{(n-1)}/(n-1)}}$$

t-Distribution

Definition

Let Z be a standard normal random variable and W be a chi-squared random variable with v degrees of freedom. If Z and W are independent, then

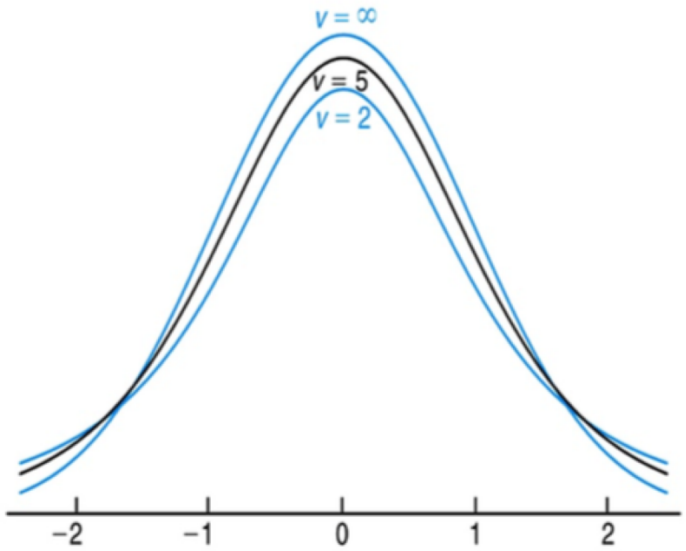
$$T = \frac{Z}{\sqrt{W/v}}$$

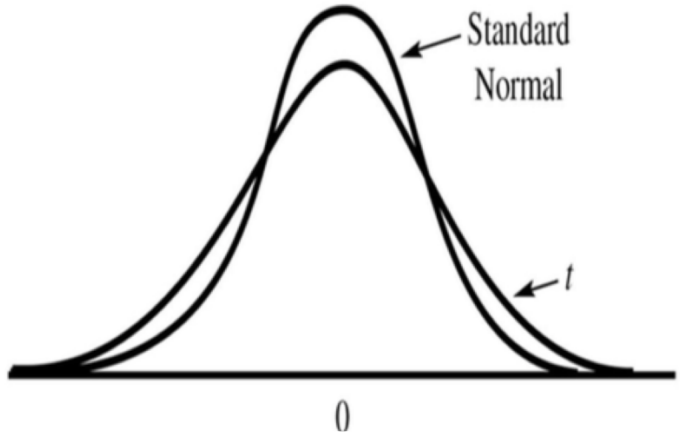
is said to have a t distribution with v df.

The density of T is (not required):

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

Properties of t-Distribution





Properties of t-Distribution

- *Each t_v curve is bell-shaped and centred at 0.*
- *Each t_v curve is spread out more than the standard normal (Z) curve.*
- *As v increases, the spread of the corresponding t_v curve decreases.*
- *As $v \to \infty$, the sequence of t_v curves approaches the standard normal curve (the Z curve is called a t curve with $df = \infty$).*

Example: # 7.30

Suppose that Z has a standard normal distribution and that Y is an independent χ^2-distributed random variable with v df. Then, according to Definition 7.2, $T = \frac{Z}{\sqrt{W/v}} \sim t_v$.

(1) If Z has a standard normal distribution, find $E(Z)$ and $E(Z^2)$.

(2) Recall that $E(W^\alpha) = \frac{\Gamma([v/2]+\alpha)}{\Gamma(v/2)} 2^\alpha$ if $v > -2\alpha$. Show that $E(T) = 0, v > 1$ and $V(T) = \frac{v}{v-2}, v > 2$.

F-Distribution

Definition

Let $W_1 \sim \chi^2_{v_1}$ and $W_2 \sim \chi^2_{v_2}$ be independent. Then

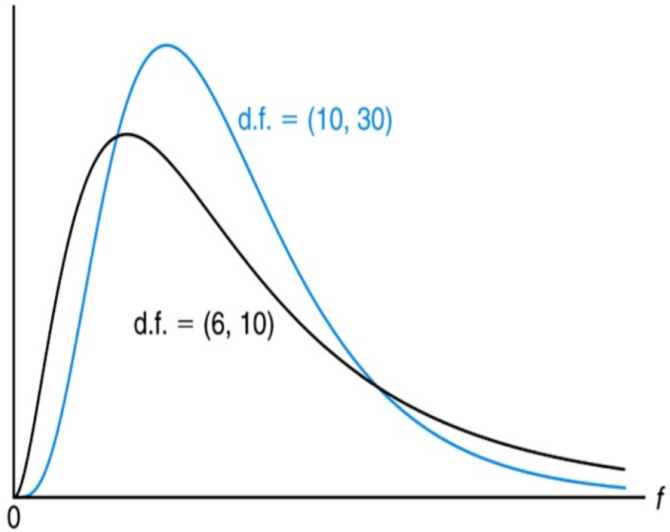
$$F = \frac{W_1/v_1}{W_2/v_2}$$

has the F-distribution with v_1 and v_2 degrees of freedom (d.f.).

- We write F_{v_1,v_2} or $F(v_1, v_2)$.
- The density of F is (not required):

$$h(f) = \begin{cases} \frac{[(v1+v2)/2](v1/v2)^{v1/2}}{\Gamma(v_1/2)\Gamma(v_2/2)} \frac{f^{(v_1/2)-1}}{(1+v_1 f/v_2)^{(v_1+v_2)/2}}, & f > 0 \\ 0 & f \leq 0 \end{cases}$$

Properties of F-Distribution



Properties of F-Distribution

Right skewed.

Table 7 page 852.

F-Distribution

The next theorem is useful to compare the variances of two normal distributions.

Theorem

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

$$\frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2} \sim F(n_1 - 1, n_2 - 1)$$

Proof:

Example

If we take independent samples of size $n_1 = 6$ and $n_2 = 10$ from two normal populations with equal population variances, find the number b such that

$$P(\frac{S_1^2}{S_2^2} \leq b) = 0.95.$$

Example: # 7.34

Show that

- $E(F) = \frac{v_2}{v_2-2}, v_2 > 2.$
- $V(F) = \frac{2v_2^2(v_1+v_2-2)}{v_1(v_2-2)^2(v_2-4)}, v_2 > 4.$

The Central Limit Theorem

Central Limit Theorem (CLT)

Let $Y_1, Y_2, \ldots, Y_n$ be a random sample from a population with mean μ and variance σ^2, but unknown (or non-normal) distribution. Then if n is sufficiently large, $\bar{Y}$ is approximately normally distributed with mean μ and variance σ^2/n, i.e.
$\bar{Y} \approx N(\mu, \sigma^2/n)$.

- The Central Limit Theorem (CLT) states that the sample mean from any probability distribution (as long as they have a mean and variance) will have an approximate normal distribution, if the sample is sufficiently large.
- "Large n" means $n \geq 30$ in general, but in some cases may even be much less.
- The larger the sample size, the more nearly normally distributed is the population of all possible sample means.
- For fairly symmetric distributions, $n > 15$ will be sufficient.

Example: # 7.42

The fracture strength of tempered glass averages 14 (measured in thousands of pounds per square inch) and has standard deviation 2.

(1) What is the probability that the average fracture strength of 100 randomly selected pieces of this glass exceeds 14.5?

(2) Find an interval that includes, with probability 0.95, the average fracture strength of 100 randomly selected pieces of this glass.

Normal Approximation to the Binomial

The normal distribution can be used to approximate binomial probabilities when there is a very large number of trials and when np and $n(1-p)$ are both large ($np \geq 10$ and $n(1-p) \geq 10$).

Normal Approximation to the Binomial

If $Y \sim Bin(n, p)$, *then* $Y \approx N(\mu = np, \sigma^2 = npq)$.

To improve the accuracy of the approximation, we usually use a correction factor, called continuity correction, to take into account that the binomial random variable is discrete while the normal is continuous.

Normal Approximation to the Binomial

The basic idea is to treat the discrete value k as the continuous interval from $k - 0.5$ to $k + 0.5$ giving the following adjustments:

- $P(Y = k) = P(k - 0.5 \leq Y \leq k + 0.5)$
- $P(Y \leq k) = P(Y \leq k + 0.5)$
- $P(k \leq Y) = P(k - 0.5 \leq Y)$

Example # 7.80

The median age of residents of the united states is 31 years. If a survey of 100 randomly selected U.S. residents is to be taken, what is the approximate probability that at least 60 will be under 31 years of age?