Chapter 7: Sampling Distributions

Juhyung Lee

Department of Statistics
University of Florida

Sampling Distributions

Recall that we defined a statistic to be a numerical summary of a sample taken from the population. Here is a more rigorous definition.

Definition (Statistic)

A function of one or more random variables that does not depend upon any unknown parameter is called a **statistic**. A statistic is a random variable.

Example

- The random variable $Y = \sum_{i=1}^{n} X_i$ is a statistic.
- The random variable $Z = (X_1 - \mu)/\sigma$ is not a statistic unless μ and σ are known numbers.

Definition (Sampling Distribution)

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

Sampling Distribution of Means and the Central Limit Theorem

Theorem

Suppose $X_1, X_2, \ldots, X_n$ *are independent and identically distributed (iid) with* $E(X) = \mu$ *and* $Var(X) = \sigma^2 < \infty$. *Let* $\overline{X} = n^{-1}\sum_{i=1}^{n} X_i$ *denote the sample mean. Then*

$$E(\overline{X}) = \mu \text{ and } Var(\overline{X}) = \frac{\sigma^2}{n}.$$

Theorem

Suppose $X_1, X_2, \ldots, X_n$ *are iid* $N(\mu, \sigma^2)$. *Then*

$$\overline{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right).$$

Sampling Distribution of Means and the Central Limit Theorem

Theorem (Theorem 8.2 of WMMY: Central Limit Theorem (CLT))

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 the limiting form of the distribution of

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

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

The sample size $n = 30$ is a guideline to use for the CLT.

Sampling Distribution of Means and the Central Limit Theorem

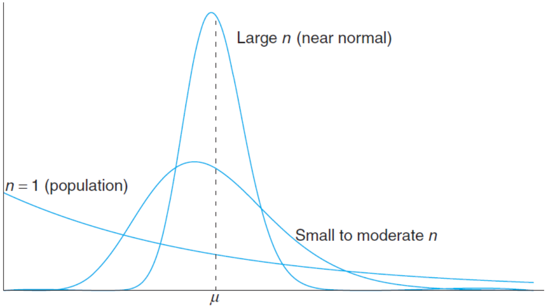

Figure 8.1: Illustration of the Central Limit Theorem (distribution of $\bar{X}$ for $n = 1$, moderate n, and large n).

Example

An electrical firm manufactures light bulbs that have a length of life that is distributed with mean equal to 800 hours and a standard deviation of 40 hours. Find the probability that a random sample of 64 bulbs will have an average life of less than 790 hours.

Sampling Distribution of Means and the Central Limit Theorem

The CLT can be easily extended to the two-sample, two-population case.

Theorem (Theorem 8.3 of WMMY)

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

$$E(\overline{X}_1 - \overline{X}_2) = \mu_1 - \mu_2 \text{ and } Var(\overline{X}_1 - \overline{X}_2) = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}.$$

Hence,

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

The normal approximation is usually good if $n_1 \geq 30$ and $n_2 \geq 30$.

Sampling Distribution of Means and the Central Limit Theorem

Example (Example 8.6 of WMMY)

The television picture tubes of manufacturer A have a mean lifetime of 6.5 years and a standard deviation of 0.9 year, while those of manufacturer B have a mean lifetime of 6.0 years and a standard deviation of 0.8 year. What is the probability that a random sample of 36 tubes from manufacturer A will have a mean lifetime that is at least 1 year more than the mean lifetime of a sample of 49 tubes from manufacturer B?

Sampling Distribution of S^2

- The sampling distribution of $\overline{X}$ is used to learn about the population mean μ.
- Similarly, the sampling distribution of S^2 is used to learn about the population variance σ^2.

Theorem (Theorem 8.4 of WMMY)

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

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

t-Distribution

- In practice, a direct use of the CLT is often restricted due to lack of knowledge on the population variance σ^2.
- Then the unknown σ^2 is estimated by the sample variance S^2 and the t-distribution arises.

Theorem (Theorem 8.5 of WMMY)

Suppose $Z \sim N(0,1)$, $V \sim \chi^2_\nu$, and $Z \perp\!\!\!\perp V$ (i.e., Z and V are independent). Then

$$\frac{Z}{\sqrt{V/\nu}} \sim t_\nu,$$

*a **t-distribution** with ν degrees of freedom.*

Theorem (Corollary 8.1 of WMMY)

Suppose $X_i \overset{iid}{\sim} N(\mu, \sigma^2), i = 1, \ldots, n$. Then

$$T \triangleq \frac{\overline{X} - \mu}{S/\sqrt{n}} \sim t_{n-1}.$$

t-Distribution

- Both the standard normal distribution and the t-distribution are bell-shaped, symmetric about zero, but the t-distribution is more variable and it has heavier tails (i.e., more likely to have large/small values).
- As $\nu \to \infty$, $t_\nu \to N(0, 1)$.

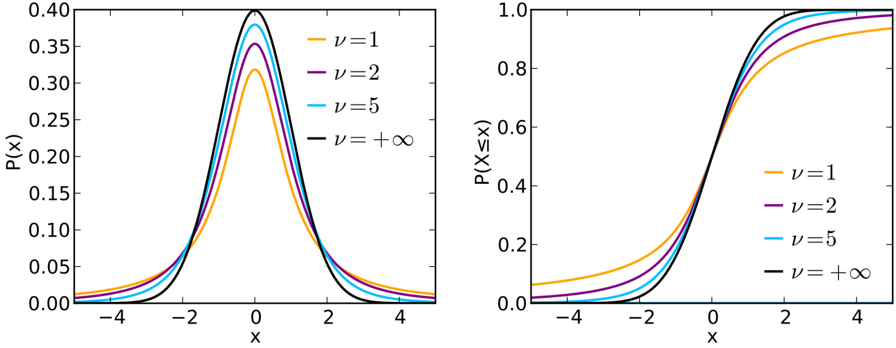

Figure: t pdfs and cdfs.

t-Distribution

t-Distribution in R

In R,

- the function `dt()` computes the t pdf;
- the function `pt()` computes the t cdf;
- the function `qt()` computes the t quantiles.

t-Distribution

Example

The t-value with $\nu = 14$ df that leaves an area of 0.025 to the right, and therefore an area of 0.975 to the left, is

$$t_{0.025,14} = -t_{0.975,14} = 2.145.$$

Note that $t_{0.025,14} = 2.145 > 1.96 = z_{0.025}$ as the t_{14} distribution has heavier tails than the standard normal distribution.

```
> qt(0.975, df = 14) # upper 0.025 quantile of t_14
[1] 2.144787
> qnorm(0.975) # upper 0.025 quantile of N(0, 1)
[1] 1.959964
>
> # probabilities to the left and right of 2.145 for t_14
> pt(2.145, df = 14)
[1] 0.9750099
> pt(2.145, 14, lower.tail = FALSE)
[1] 0.02499008
```

F-Distribution

Theorem (Theorem 8.6 of WMMY)

Suppose $U \sim \chi^2_{\nu_1}$, $V \sim \chi^2_{\nu_2}$, and $U \perp\!\!\!\perp V$. Then

$$\frac{U/\nu_1}{V/\nu_2} \sim F_{\nu_1,\nu_2},$$

*an **F-distribution** with ν_1 and ν_2 degrees of freedom. Here, ν_1 is the numerator df and ν_2 is the denominator df.*

Theorem (Theorem 8.8 of WMMY)

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

$$F \triangleq \frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2} \sim F_{n_1-1,n_2-1}.$$

F-Distribution

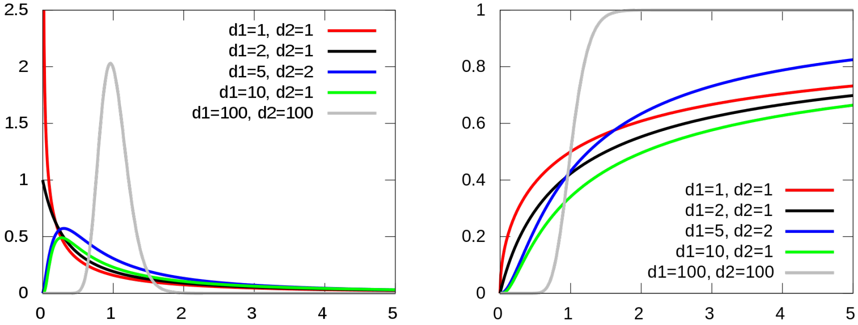

Figure: F pdfs and cdfs.

F-Distribution in R

In R,

- the function `df()` computes the F pdf;
- the function `pf()` computes the F cdf;
- the function `qf()` computes the F quantiles.

F-Distribution

Example

The F-value with 6 and 10 df, leaving an area of 0.05 to the right, is

$$F_{0.05,6,10} = 3.217.$$

```
> qf(0.95, df1 = 6, df2 = 10) # upper 0.05 quantile of F_{6,10}
[1] 3.217175
>
> # probabilities to the left and right of 3.217 for F_{6,10}
> pf(3.217, df1 = 6, df2 = 10)
[1] 0.9499925
> pf(3.217, 6, 10, lower.tail = FALSE)
[1] 0.05000754
```