MAT2377

Ali Karimnezhad

Version December 13, 2016

Comments

- These slides cover material from Chapter 4.
- In class, I may use a blackboard. I recommend reading these slides before you come to the class.
- I am planning to spend 3 lectures on this chapter.
- I am not re-writing the textbook. The reference book contains many interesting and practical examples.
- There may be some typos. The final version of the slides will be posted $after$ the chapter is finished.

Random Samples and Statistic!

- Let $X_1, X_2, \ldots, X_n$ be n independent random variables, each having the same probability density function $f(x)$.

- 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 density function as

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

- Our main purpose in selecting random samples is to elicit information about the unknown population parameters.

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

Three Measures of Central Tendency

- Sample mean: $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$.

- Median: Given that the observations in a sample are $X_1, X_2, \ldots, X_n$, arranged in increasing order of magnitude, i.e., $X_{(1)} = \min_{1\le i\le n} X_i$ and $X_{(n)} = \max_{1\le i\le n} X_i$;

$$\tilde{X} = \begin{cases} X_{(\frac{n+1}{2})} & \text{if } n \text{ is odd} \\ \frac{X_{(\frac{n}{2})}+X_{(\frac{n}{2}+1)}}{2} & \text{if } n \text{ is even} \end{cases}$$

- The sample mode is the value of the sample that occurs most often.

Examples:

- Suppose a data set consists of the following observations:

$$0.32, 0.53, 0.28, 0.37, 0.47, 0.43, 0.36, 0.42, 0.38, 0.43.$$

Compute sample mean, median and mode.

- Consider the following measurements, in liters, for two samples of orange juice bottled by companies A and B:

Sample A	0.97	1.00	0.94	1.03	1.06
Sample B	1.06	1.01	0.88	0.91	1.14

Both samples have the same mean, 1.00 liter.

We say that the variability, or the dispersion, of the observations from the average is less for sample A than for sample B. Therefore, in buying orange juice, we can feel more confident that the bottle we select will be close to the advertised average if we buy from company A.

Three Measures of Variability

- Sample variance:

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

- Sample standard deviation: $S = \sqrt{S^2}$.
- Sample range: $X_{(n)} - X_{(1)}$.

Example:

- 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.

Sampling Distributions

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

- The first important sampling distribution to be considered is that of the mean $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$.

- Suppose that a random sample of n observations is taken from a normal population with mean μ and variance σ^2.

- Each observation X_i, $i = 1, 2, \ldots, n$, of the random sample will then have the same normal distribution as the population being sampled. It can be shown that $\bar{X} \sim N(\mu_{\bar{X}}, \sigma^2_{\bar{X}})$, where

$$\mu_{\bar{X}} = E(\bar{X}) = \mu, \quad \sigma^2_{\bar{X}} = Var(\bar{X}) = \frac{\sigma^2}{n}.$$

Central Limit Theorem

If we are sampling from a population with unknown distribution, either finite or infinite, the sampling distribution of $\bar{X}$ will still be approximately normal with mean μ and variance $\frac{\sigma^2}{n}$, provided that the sample size is large.

- 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

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

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

- The normal approximation for $\bar{X}$ will generally be good if $n \geq 30$.

Examples:

- 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.

- Traveling between two campuses of a university in a city via shuttle bus takes, on average, 28 minutes with a standard deviation of 5 minutes. In a given week, a bus transported passengers 40 times.
  What is the probability that the average transport time, i.e., the average for 40 trips, was more than 30 minutes? Assume the mean time is measured to the nearest minute.

Sampling Distribution of Mean Differences

- If independent samples of size n_1 and n_2 are drawn at random from two populations 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, $\bar{X}_1 - \bar{X}_2$, is approximately normally distributed with mean and variance given by

$$\mu_{\bar{X}_1-\bar{X}_2} = E[\bar{X}_1 - \bar{X}_2] = \mu_1 - \mu_2,$$

$$\sigma^2_{\bar{X}_1-\bar{X}_2} = Var[\bar{X}_1 - \bar{X}_2] = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}.$$

Hence,

$$Z = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}},$$

is approximately a standard normal variable.

Example:

Two independent experiments are run in which two different types of paint are compared. Eighteen specimens are painted using type A, and the drying time, in hours, is recorded for each. The same is done with type B. The population standard deviations are both known to be 1.0 hour. Assuming that the mean drying time is equal for the two types of paint, find $P(\bar{X}_A - \bar{X}_B > 1.0)$, where $\bar{X}_A$ and $\bar{X}_B$ are average drying times for samples of size $n_A = n_B = 18$.

Sampling Distribution of S^2

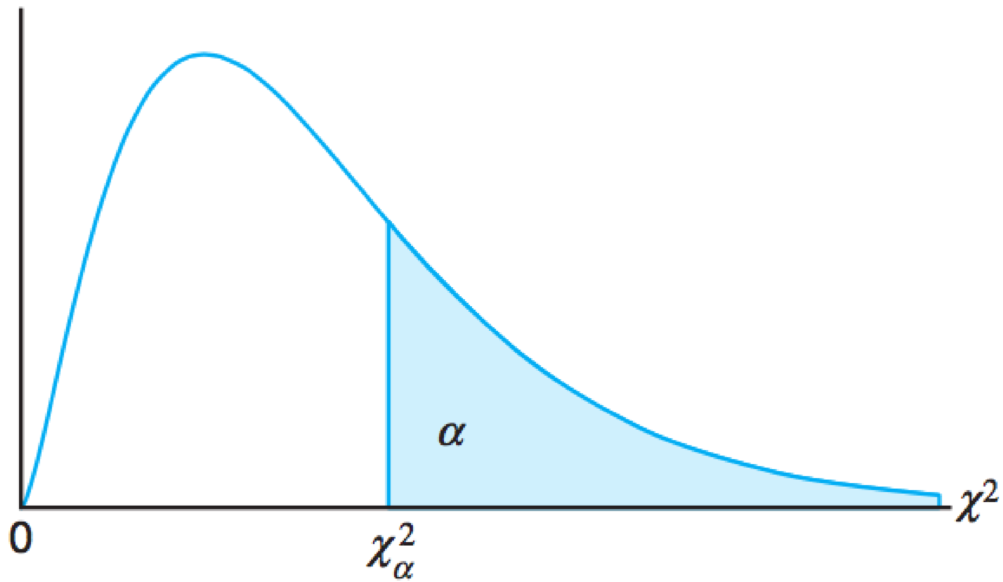
- If S^2 is the variance of a random sample of size n taken from a normal population having the variance σ^2, then the statistic

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

a chi-squared distribution with $\nu = n - 1$ degrees of freedom.

$$P\left(S^2 > k\right) = ?$$

$$P\left(\frac{(n-1)S^2}{\sigma^2} > \chi^2_\alpha\right) = \alpha$$



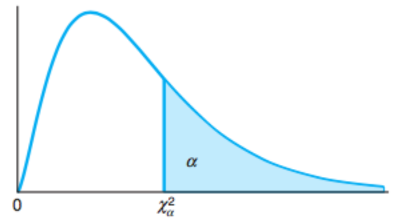


Table A.5 Critical Values of the Chi-Squared Distribution

	α									
v	**0.995**	**0.99**	**0.98**	**0.975**	**0.95**	**0.90**	**0.80**	**0.75**	**0.70**	**0.50**
1	0.0^4393	0.0^3157	0.0^3628	0.0^3982	0.00393	0.0158	0.0642	0.102	0.148	0.455
2	0.0100	0.0201	0.0404	0.0506	0.103	0.211	0.446	0.575	0.713	1.386
3	0.0717	0.115	0.185	0.216	0.352	0.584	1.005	1.213	1.424	2.366
4	0.207	0.297	0.429	0.484	0.711	1.064	1.649	1.923	2.195	3.357
5	0.412	0.554	0.752	0.831	1.145	1.610	2.343	2.675	3.000	4.351
6	0.676	0.872	1.134	1.237	1.635	2.204	3.070	3.455	3.828	5.348
7	0.989	1.239	1.564	1.690	2.167	2.833	3.822	4.255	4.671	6.346
8	1.344	1.647	2.032	2.180	2.733	3.490	4.594	5.071	5.527	7.344
9	1.735	2.088	2.532	2.700	3.325	4.168	5.380	5.899	6.393	8.343
10	2.156	2.558	3.059	3.247	3.940	4.865	6.179	6.737	7.267	9.342

t-Distribution

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

$$T = \frac{Z}{\sqrt{\frac{V}{\nu}}},$$

is given by the density function

$$f(t) = \frac{\Gamma(\frac{\nu+1}{2})}{\Gamma(\frac{\nu}{2})\sqrt{\pi\nu}}\left(1+\frac{t^2}{\nu}\right)^{-\frac{\nu+1}{2}}, \quad -\infty < t < \infty.$$

This is known as the Student t-distribution with ν degrees of freedom. We briefly write $T \sim t(\nu)$.

Main Result

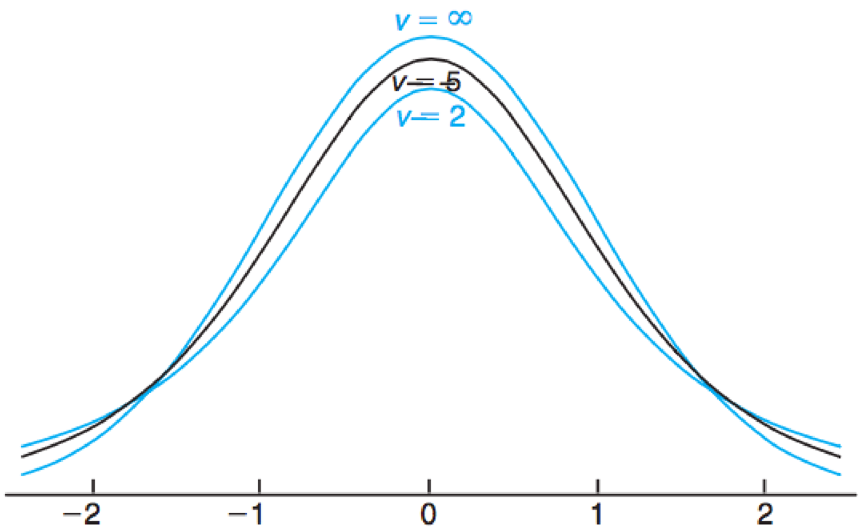
Let $X_1, X_2, \ldots, X_n$ be independent random variables that are all normal with mean μ and standard deviation σ. Let

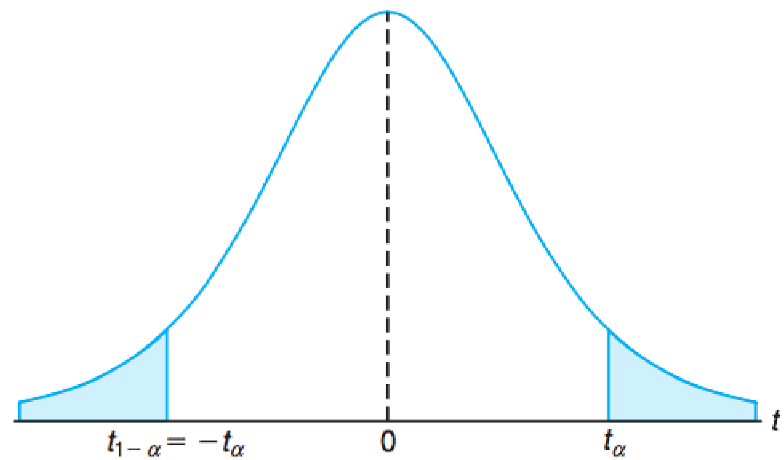
$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \quad \text{and} \quad S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

Then,

$$T = \frac{\bar{X} - \mu}{\frac{S}{\sqrt{n}}} \sim t(n-1).$$

What Does the t-Distribution Look Like?





Example:

Find k such that $P(k < T < -0.258) = 0.10$ for a random sample of size 15 selected from a normal distribution, where $T = \frac{\bar{X}-\mu}{\frac{S}{\sqrt{n}}} \sim t(n-1)$.

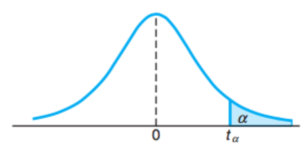


Table A.4 Critical Values of the t-Distribution

	α						
v	**0.40**	**0.30**	**0.20**	**0.15**	**0.10**	**0.05**	**0.025**
1	0.325	0.727	1.376	1.963	3.078	6.314	12.706
2	0.289	0.617	1.061	1.386	1.886	2.920	4.303
3	0.277	0.584	0.978	1.250	1.638	2.353	3.182
4	0.271	0.569	0.941	1.190	1.533	2.132	2.776
5	0.267	0.559	0.920	1.156	1.476	2.015	2.571
6	0.265	0.553	0.906	1.134	1.440	1.943	2.447
7	0.263	0.549	0.896	1.119	1.415	1.895	2.365
8	0.262	0.546	0.889	1.108	1.397	1.860	2.306
9	0.261	0.543	0.883	1.100	1.383	1.833	2.262
10	0.260	0.542	0.879	1.093	1.372	1.812	2.228
11	0.260	0.540	0.876	1.088	1.363	1.796	2.201
12	0.259	0.539	0.873	1.083	1.356	1.782	2.179
13	0.259	0.538	0.870	1.079	1.350	1.771	2.160
14	0.258	0.537	0.868	1.076	1.345	1.761	2.145
15	0.258	0.536	0.866	1.074	1.341	1.753	2.131

F-Distribution

- Let U and V be two independent random variables having chi-squared distributions with ν_1 and ν_2 degrees of freedom, respectively.

  Then the distribution of the random variable $X = \frac{\frac{U}{\nu_1}}{\frac{V}{\nu_2}}$ is given by the density function

$$f(x) = \frac{\Gamma(\frac{\nu_1+\nu_2}{2})\left(\frac{\nu_1}{\nu_2}\right)^{\frac{\nu_1}{2}}}{\Gamma(\frac{\nu_1}{2})\Gamma(\frac{\nu_2}{2})} \frac{x^{\frac{\nu_1}{2}-1}}{\left(1+\frac{\nu_1}{\nu_2}x\right)^{\frac{1}{2}(\nu_1+\nu_2)}}, \quad x > 0.$$

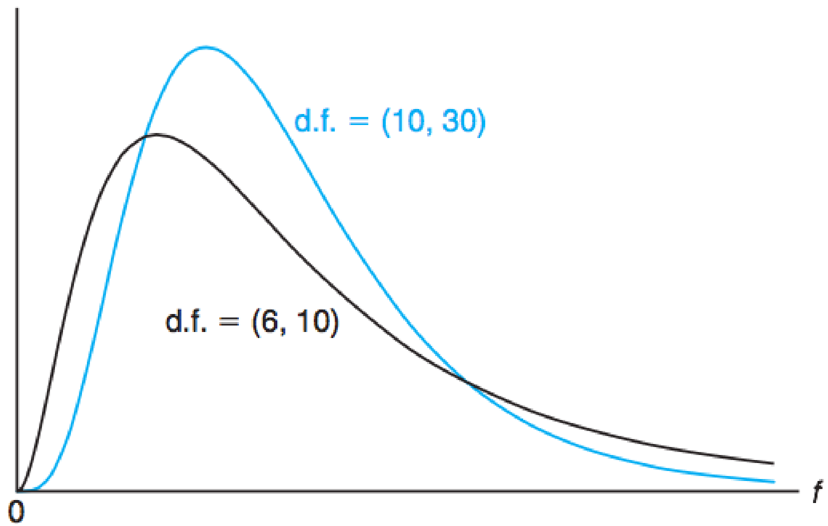
This is known as the F-distribution with ν_1 and ν_2 degrees of freedom. We briefly write $X \sim F(\nu_1, \nu_2)$.

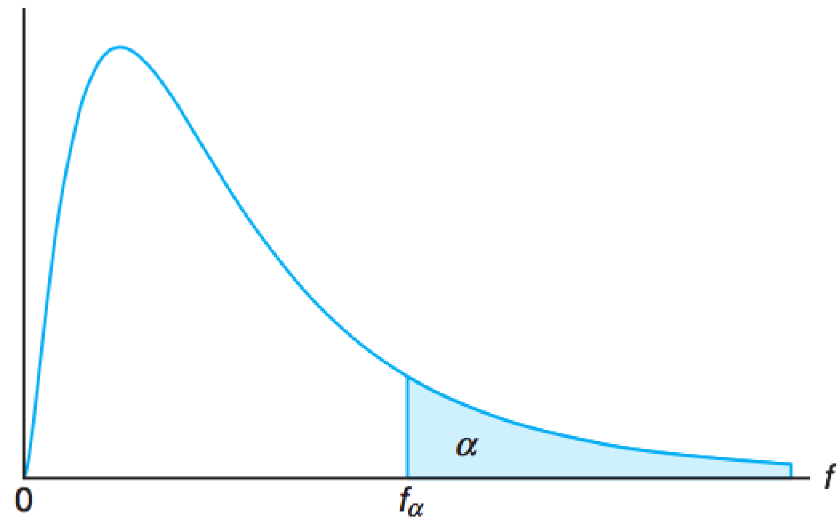
Main Result

Suppose that random samples of size n_1 and n_2 are selected from two normal populations with variances σ_1^2 and σ_2^2, respectively. Then,

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

What Does the F-Distribution Look Like?





Examples:

Find c such that $P(X > c) = 0.05$, where $X \sim F(9, 12)$.

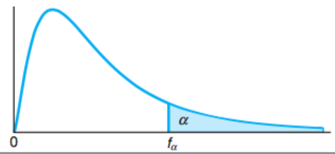


Table A.6 Critical Values of the *F*-Distribution

	$f_{0.05}(v_1, v_2)$								
	v_1								
v_2	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
1	161.448	199.500	215.707	224.583	230.162	233.986	236.768	238.883	240.5433
2	18.513	19.000	19.164	19.247	19.296	19.330	19.353	19.371	19.385
3	10.128	9.552	9.277	9.117	9.013	8.941	8.887	8.845	8.812
4	7.709	6.944	6.591	6.388	6.256	6.163	6.094	6.041	5.999
5	6.608	5.786	5.409	5.192	5.050	4.950	4.876	4.818	4.772
6	5.987	5.143	4.757	4.534	4.387	4.284	4.207	4.147	4.099
7	5.591	4.737	4.347	4.120	3.972	3.866	3.787	3.725	3.686
8	5.318	4.459	4.066	3.838	3.687	3.581	3.500	3.438	3.388
9	5.117	4.256	3.863	3.633	3.482	3.373	3.293	3.230	3.179
10	4.964	4.103	3.708	3.478	3.326	3.217	3.135	3.072	3.020
11	4.844	3.982	3.587	3.357	3.204	3.095	3.012	2.948	2.896
12	4.747	3.885	3.490	3.259	3.106	2.996	2.913	2.849	2.796
13	4.667	3.806	3.411	3.179	3.025	2.915	2.832	2.767	2.714
14	4.600	3.739	3.344	3.112	2.958	2.848	2.764	2.699	2.646
15	4.543	3.682	3.287	3.056	2.901	2.790	2.707	2.641	2.588