TRAINING COURSES ON APPLIED ECONOMETRIC ANALYSIS (SUMMER SCHOOL) FOR YOUNG ECONOMISTS / RESEARCHERS ORGANIZED BY WIUT AND IFPRI JUNE 4-15, 2018

Review of Probability

Farrukh Ataev
Lecturer, WIUT
6 June 2018

Outline

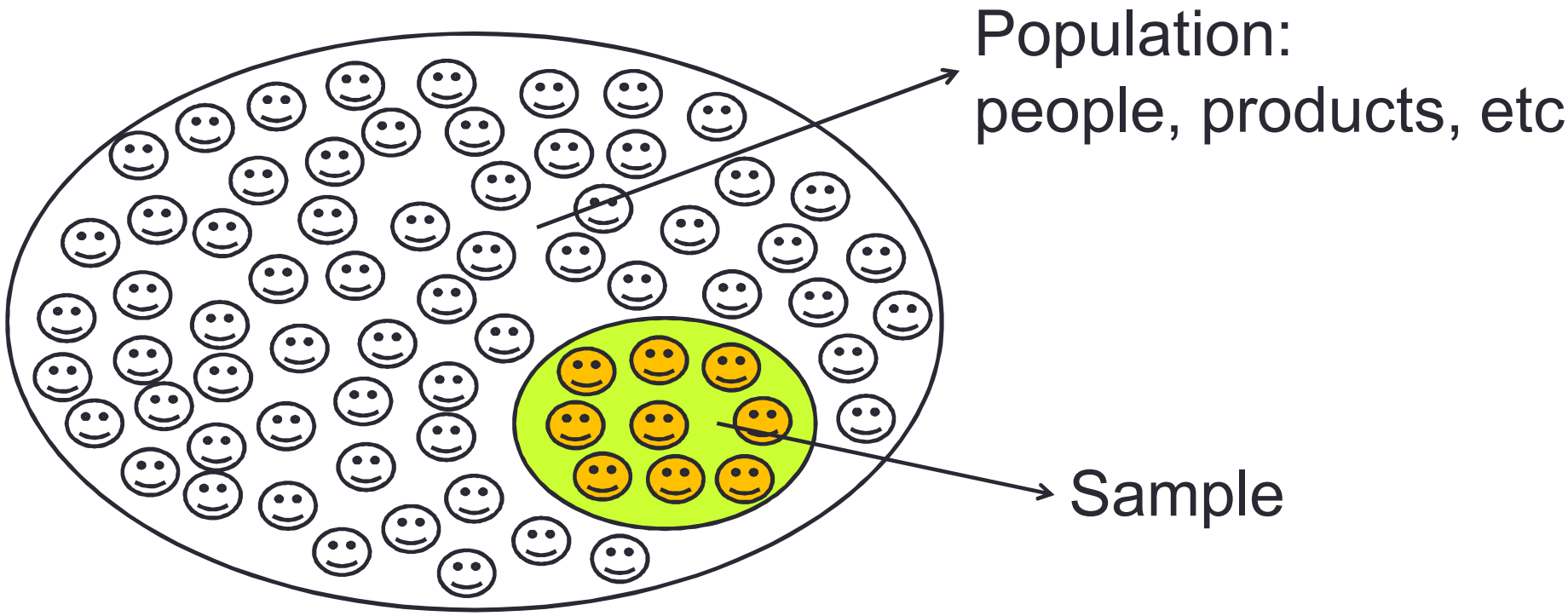
- **Session 1 Review of probability**
  - (Tue 5 June at 13:30-15:00)
- **Session 2 Review of probability (continued)**
  - (Tue 5 June at 15:30-17:00)
- **Session 3 Probability distributions**
  - (Wed 6 June at 9:00-10:30)
- **Session 4 Probability distributions (continued)**
  - (Wed 6 June at 11:00-12:30)

Session 4 outline

- Sampling distribution and Central Limit Theorem
- Confidence interval
- F distribution
- Chi-squared distribution

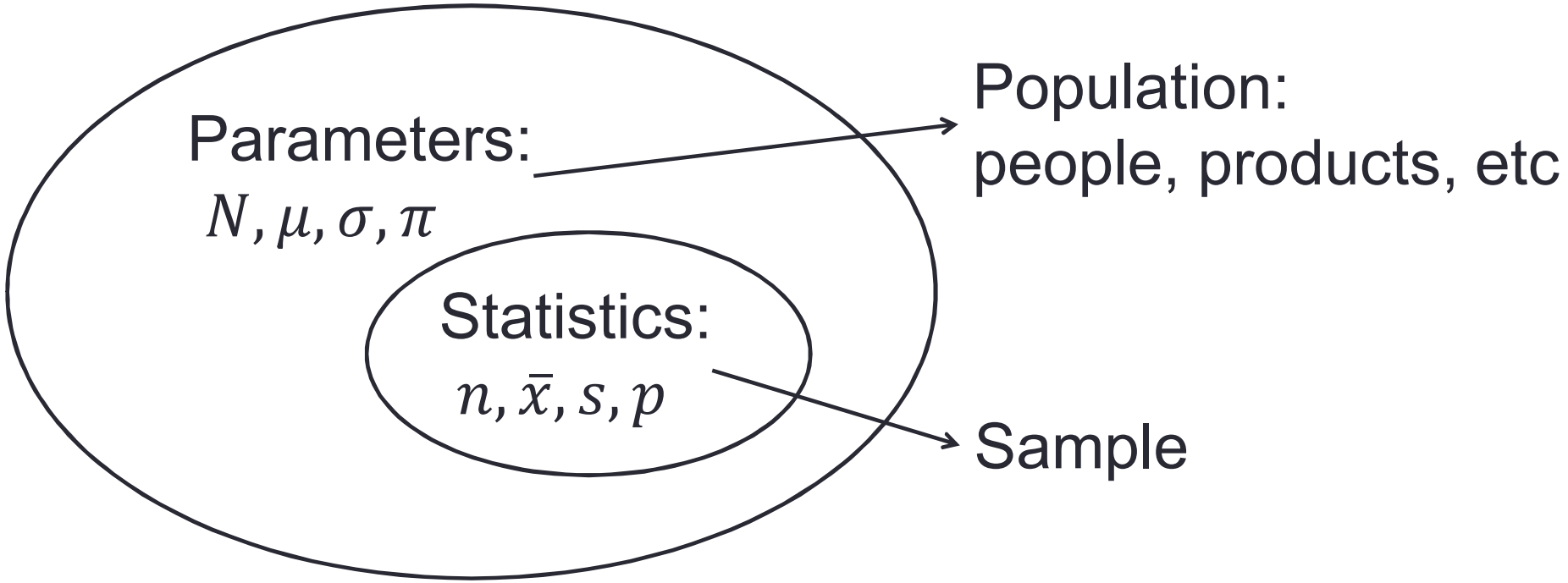
Statistics

- Science of collecting, organizing, presenting and interpreting data
- Population is a collection of all elements in a study
- Sample is a subset of the population



Statistics

- Descriptive (tabular, graphical and numerical methods to summarize data)
- Inferential (using sample data to make general conclusions (inferences) about population)

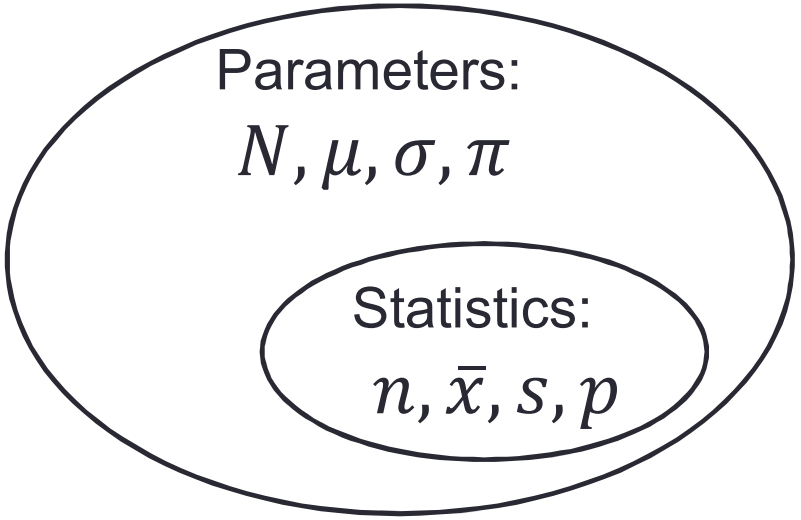


Statistics example

In a class of 50 students, 20 are foreigners. The teacher held a questionnaire on study hours of a sample of 36 students (9 are foreigners) and obtained the following data:

3.2	**2.5**	**2.7**	**5.1**	**4.1**	**3.5**	**2.8**	**2.5**	**3.3**	**3.9**	**4.4**	**3.6**
3.0	**2.9**	**2.6**	**3.1**	**3.2**	**3.4**	**2.4**	**2.5**	**2.8**	**2.7**	**3.8**	**6.2**
4.3	**1.5**	**1.8**	**2.8**	**3.2**	**3.3**	**3.7**	**2.9**	**2.8**	**4.8**	**5.2**	**3.0**

From past research the teacher knows the standard deviation is 1.1.



- What is the size of the population (or sample)?
- What is the proportion of foreigners in population (or sample)?
- What is the population mean (or SD) study time? (Descriptive statistics)
- What is the sample mean (or SD) study time?
- Can the sample mean time be generalized to the whole class? (Inferential statistics)

Sampling criteria

- Representative of population
- The goal is to use the results from sample to estimate the population, that is, the results of the sample are generalized to the population.

Sampling methods

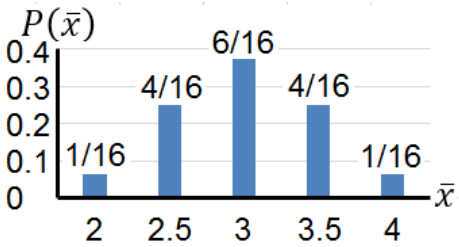
- Probability (random) (every unit of population has a certain probability of being selected)
  - **Simple random** (number all and select from a hat)
  - **Systematic** (select a starting point and every *k*th, sample size is n = N/k)
  - **Stratified** (divide into strata (groups), randomly select proportionally from **each** stratum and combine them to form a full sample, each stratum is homogenous) (divide students into strata by gender, age, level)
  - **Cluster** (divide into clusters (groups), randomly select **some** groups and all their members, each cluster is heterogenous)
- Non-probability (non-random) (some elements of population have no chance of selection, thus could be subject to bias)
  - **Convenience** (select from population easier to reach) (select the first n people who enter a store, people in a street, volunteers for a study)
  - **Quota** (divide into strata, non-randomly select proportionally from each stratum and combine them to form a full sample)
  - **Purposive/Judgemental** (select based on an expert's opinion)

Sampling distribution

- Sampling distribution is defined as the probability distribution of sample statistic.
- The sampling distribution has a mean and standard deviation.
- The standard deviation of the sampling distribution is called as standard error.

Example: A team consists of 4 students, who study 3, 2, 4, and 3 hours. Find all possible random samples with replacement of size 2 and compute the mean and standard deviation of the sampling distribution.

x	3 3	3 2	3 4	3 3	2 3	2 2	2 4	2 3	4 3	4 2	4 4	4 3	3 3	3 2	3 4	3 3
$\bar{x}$	3	2.5	3.5	3	2.5	2	3	2.5	3.5	3	4	3.5	3	2.5	3.5	3



$\bar{x}$	2	2.5	3	3.5	4		Sampling distribution	Population	CLT
$P(\bar{x})$	1/16	4/16	6/16	4/16	1/16				
$\bar{x} \cdot P(\bar{x})$	1/8	5/8	9/8	7/8	2/8	3	$E(\bar{x}) = \mu_{\bar{x}} = 3$	$\mu = 3$	$\mu_{\bar{x}} = \mu$
$\bar{x}^2 \cdot P(\bar{x})$	0.25	1.5625	3.375	3.0625	1	9.25	$\sigma_{\bar{x}} = \sqrt{9.25 - 3^2} = 0.5$	$\sigma = 0.71$	$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{2}}$

Central Limit Theorem (for mean)

If a random sample of size *n* is drawn from a population with mean μ and standard deviation σ, the distribution of the sample mean $\bar{x}$ approaches a normal distribution with mean μ and standard deviation $\sigma/\sqrt{n}$ as the sample size increases.

2 properties:

1) If the population is **normal**, the sample mean has a normal distribution centered at μ, with a standard error equal to $\sigma/\sqrt{n}$.

2) If the population is **not normal**, the sample mean has more normal distribution centered at μ, with a standard error equal to $\sigma/\sqrt{n}$ as the sample size increases (n≥30).

Confidence interval

- The interval (margin) at which the population parameter is located around the sample statistic.
- The larger the sample size the more precise the estimate.

Example

In a class of 50 students, 20 are foreigners. The teacher held a questionnaire on study hours of a sample of 36 students (9 are foreigners) and obtained the following data:

3.2	**2.5**	**2.7**	**5.1**	**4.1**	**3.5**	**2.8**	**2.5**	**3.3**	**3.9**	**4.4**	**3.6**
3.0	**2.9**	**2.6**	**3.1**	**3.2**	**3.4**	**2.4**	**2.5**	**2.8**	**2.7**	**3.8**	**6.2**
4.3	**1.5**	**1.8**	**2.8**	**3.2**	**3.3**	**3.7**	**2.9**	**2.8**	**4.8**	**5.2**	**3.0**

From past research the teacher knows the standard deviation is 1.1.

Find a) 90% b) 95% confidence interval of the population mean.

Population:

$(N = 50, \mu = ?, \sigma = 1.1, \pi = 0.4)$

Sample:

$(n = 36, \bar{x} = 3.3, s = 1, p = 0.25)$

Confidence interval: $\bar{x} - z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}} \leq \mu \leq \bar{x} + z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}}$

$$3.3 - ME \leq \mu \leq 3.3 + ME$$

Note: $\sqrt{\frac{N-n}{N-1}}$ is a finite population correction factor.

Example (cont.)

Find a) 90% b) 95% conf. interval of the population mean.

$$\bar{x} - z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}} \leq \mu \leq \bar{x} + z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}}$$

Solution: $\bar{x} = 3.3, \sigma = 1.1, N = 50, n = 36.$

a) $1 - \alpha = 0.9 \Rightarrow \alpha = 0.1 \Rightarrow \alpha/2 = 0.05.$

$z_{0.05} = NORM.S.INV(0.05) = -1.645.$

$$3.3 - 1.645 \cdot \frac{1.1}{6} \sqrt{\frac{50-36}{50-1}} \leq \mu \leq 3.3 + 1.645 \cdot \frac{1.1}{6} \sqrt{\frac{50-36}{50-1}} \Rightarrow$$

$$3.14 \leq \mu \leq 3.46$$

b) $1 - \alpha = 0.95 \Rightarrow \alpha = 0.05 \Rightarrow \alpha/2 = 0.025.$

$z_{0.025} = NORM.S.INV(0.025) = -1.96.$

$$3.3 - 1.96 \cdot \frac{1.1}{6} \sqrt{\frac{50-36}{50-1}} \leq \mu \leq 3.3 + 1.96 \cdot \frac{1.1}{6} \sqrt{\frac{50-36}{50-1}} \Rightarrow$$

$$3.11 \leq \mu \leq 3.49$$

Exercise

We would like to estimate the mean amount of cash carried by executives in the hotel industry. We have a sample of size n=36, drawn from a normally distributed population with a known standard deviation $\sigma = 4$. The sample mean is 72.

a) For a confidence level of 95%, calculate the margin of error.

b) Calculate the width and the limits of the confidence interval for the unknown population mean.

c) Interpret the interval.

Exercise solution

$$\bar{x} - z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{x} + z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}$$

Answer: $\bar{x} = 72, \sigma = 4, n = 36$

a) For a conf. level of 95%, calculate the margin of error.

$$1 - \alpha = 0.95 \Rightarrow \alpha = 0.05 \Rightarrow \alpha/2 = 0.025.$$

$$z_{0.025} = NORM.S.INV(0.025) = -1.96.$$

$$z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}} = 1.96 \cdot \frac{4}{\sqrt{36}} = 1.31.$$

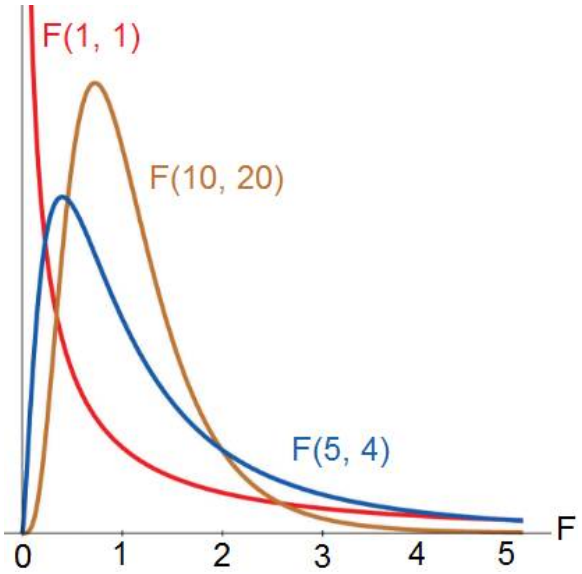
b) Calculate the width and the limits of the confidence interval for the unknown population mean.

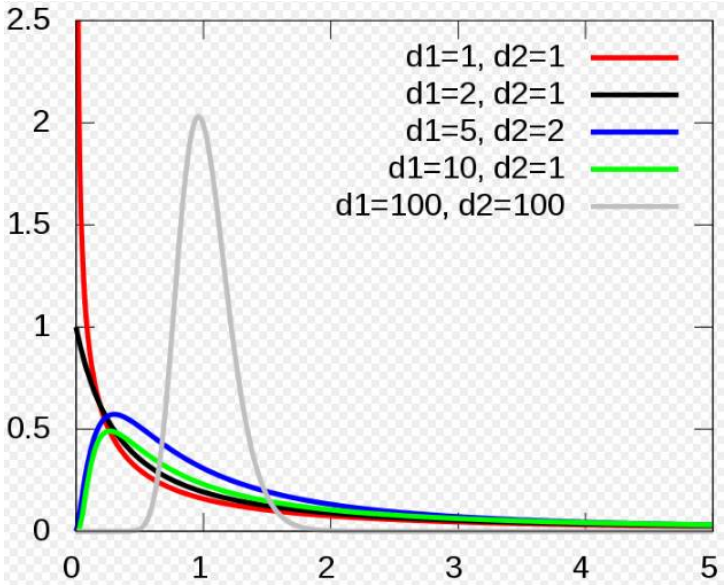
$$72 - 1.31 \leq \mu \leq 72 + 1.31 \Rightarrow$$
$$70.69 \leq \mu \leq 73.31$$

c) Interpret the interval.

Continuous: F-distribution $F(d_1, d_2)$

A continuous statistical distribution which arises in the testing of whether two observed samples have the same variance or ANOVA test (whether the means of several groups are equal).

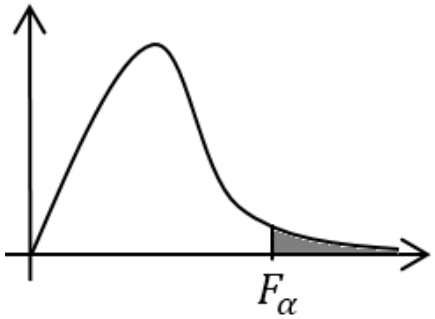




Exercise: Is the F-curve symmetric? As d_1 and d_2 get bigger, what curve does the F-curve approach?

F-distribution table $F(d_1, d_2)$

d_2/d_1	p	1	2	3	4	5	6	7	8	9
1	0.100	39.86	49.50	53.59	55.83	57.24	58.20	58.91	59.44	59.86
	0.050	161.45	199.50	215.71	224.58	230.16	233.99	236.77	238.88	240.54
	0.025	647.79	799.50	864.16	899.58	921.85	937.11	948.22	956.66	963.28
	0.010	4052.2	4999.5	5403.4	5624.6	5763.6	5859.0	5928.4	5981.1	6022.5
	0.001	405284	500000	540379	562500	576405	585937	592873	598144	602284
2	0.100	8.53	9.00	9.16	9.24	9.29	9.33	9.35	9.37	9.38
	0.050	18.51	19.00	19.16	19.25	19.30	19.33	19.35	19.37	19.38
	0.025	38.51	39.00	39.17	39.25	39.30	39.33	39.36	39.37	39.39
	0.010	98.50	99.00	99.17	99.25	99.30	99.33	99.36	99.37	99.39
	0.001	998.50	999.00	999.17	999.25	999.30	999.33	999.36	999.37	999.39
3	0.100	5.54	5.46	5.39	5.34	5.31	5.28	5.27	5.25	5.24
	0.050	10.13	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81
	0.025	17.44	16.04	15.44	15.10	14.88	14.73	14.62	14.54	14.47
	0.010	34.12	30.82	29.46	28.71	28.24	27.91	27.67	27.49	27.35
	0.001	167.03	148.50	141.11	137.10	134.58	132.85	131.58	130.62	129.86
4	0.100	4.54	4.32	4.19	4.11	4.05	4.01	3.98	3.95	3.94
	0.050	7.71	6.94	**6.59**	6.39	6.26	6.16	6.09	6.04	6.00
	0.025	12.22	10.65	9.98	9.60	9.36	9.20	9.07	8.90	8.90
	0.010	21.20	18.00	16.69	15.98	15.52	15.21	14.98	14.80	14.66
	0.001	74.14	61.25	56.18	53.44	51.71	50.53	49.66	49.00	48.47
5	0.100	4.06	3.78	3.62	3.52	3.45	3.40	3.37	3.34	3.32



The entries in this table show the area to the right tail of the F-curve.

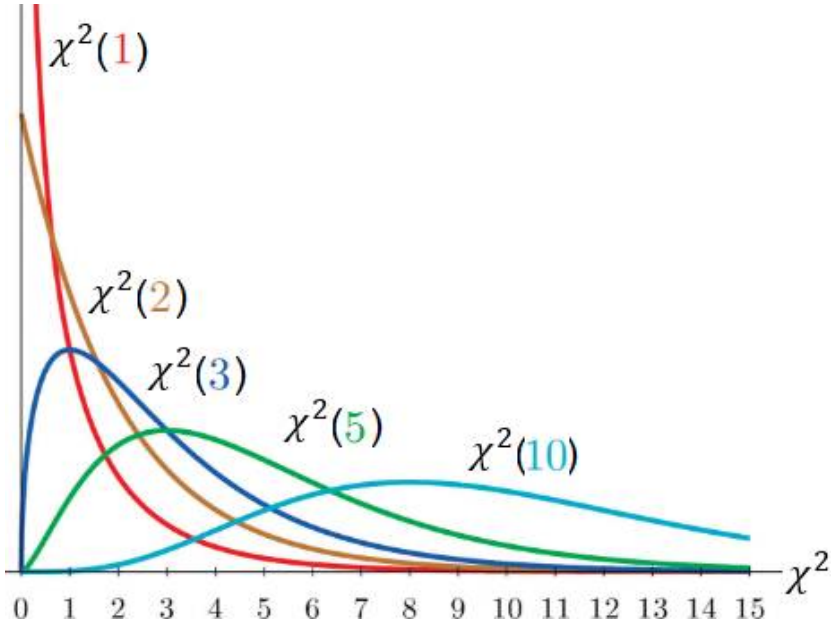
$$F_{0.05;3,4} = 6.59.$$

$$P(6.59 \leq F_{(3;4)}) = 0.05$$

Exercise: Find: a) $F_{0.01;9;2}$; b) $F_{0.01;1;5}$.

Continuous: Chi-squared distribution $\chi^2(k)$

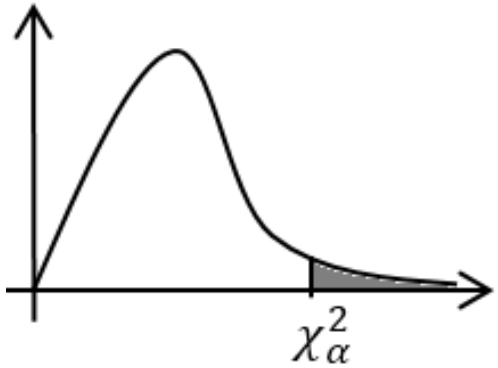
A continuous statistical distribution which arises in the testing of independence of contingency tables.



Exercise: Is the Chi-squared curve symmetric? As k gets bigger, what curve does the Chi-squared curve approach?

Chi-squared distribution table $\chi^2(k)$

k	0.995	0.990	0.975	0.950	0.900	0.100	0.050	0.025	0.010	0.005
1	0.000	0.000	0.001	0.004	0.016	2.706	3.841	5.024	6.635	7.879
2	0.010	0.020	0.051	0.103	0.211	4.605	5.991	7.378	9.210	10.597
3	0.072	0.115	0.216	0.352	0.584	6.251	7.815	9.348	11.345	12.838
4	0.207	0.297	0.484	0.711	1.064	7.779	9.488	11.143	13.277	14.860
5	0.412	0.554	0.831	1.145	1.610	9.236	11.070	12.832	15.086	16.750
6	0.676	0.872	1.237	1.635	2.204	10.645	12.592	14.449	16.812	18.548
7	0.989	1.239	1.690	2.167	2.833	12.017	14.067	16.013	18.475	20.278
8	1.344	1.646	2.180	2.733	3.490	13.362	15.507	17.535	20.090	21.955
9	1.735	2.088	2.700	3.325	4.168	14.684	16.919	19.023	21.666	23.589
10	2.156	2.558	3.247	3.940	4.865	15.987	18.307	20.483	23.209	25.188
11	2.603	3.053	3.816	4.575	5.578	17.275	19.675	21.920	24.725	26.757
12	3.074	3.571	4.404	5.226	6.304	18.549	21.026	23.337	26.217	28.300
13	3.565	4.107	5.009	5.892	7.042	19.812	22.362	24.736	27.688	29.819
14	4.075	4.660	5.629	6.571	7.790	21.064	23.685	26.119	29.141	31.319
15	4.601	5.229	6.262	7.261	8.547	22.307	24.996	27.488	30.578	32.801
16	5.142	5.812	6.908	7.962	9.312	23.542	26.296	28.845	32.000	34.267
17	5.697	6.408	7.564	8.672	10.085	24.769	27.587	30.191	33.409	35.718
18	6.265	7.015	8.231	9.390	10.865	25.989	**28.869**	31.526	34.805	37.156
19	6.844	7.633	8.907	10.117	11.651	27.204	30.144	32.852	36.191	38.582
20	7.434	8.260	9.591	10.851	12.443	28.412	31.410	34.170	37.566	39.997
21	8.034	8.897	10.283	11.591	13.240	29.615	32.671	35.479	38.932	41.401



The entries in this table show the area to the right tail of the χ^2-curve.

$$\chi^2{}_{0.05;18} = 28.869.$$

$$P\left(28.869 \le \chi^2_{(18)}\right) = 0.05$$

Exercise: Find: a) $\chi^2{}_{0.01;9}$; b) $\chi^2{}_{0.025;20}$.

Reading

1) Murray R. Spiegel, *Schaum's outline of Theory and Problems of Probability and Statistics*, McGraw-Hill, 23 edition, 1998.

2) Nitis Mukhopadhyay, Probability and Statistical Inference, Marcel Dekker, Inc. 2000.