Auxiliary distributions. Point estimators

Probability theory. Statistics

Anton Afanasev

Higher School of Economics

DSBA 211

November 19, 2022

Seminar Overview

1. Quiz
2. Auxiliary distributions
   - χ^2-distribution
   - t-distribution
   - F-distribution
3. Random variables convergence
   - Definitions
   - Law of large numbers
4. Point estimators. Specific cases
   - Sample mean

Quiz

When a fair coin is flipped 10 times, what is the chance of getting 7 or more heads? Answer in three ways.

1. Exactly, using binomial distribution.
2. Approximately, using the normal distribution.
3. Approximately, using the normal approximation with continuity correction.

Note that the normal approximation with continuity correction can be an excellent approximation, even for n as small as 10.

χ^2-distribution

- Let $Z_1, \ldots, Z_k$ be i.i.d. with $Z_i \sim \mathcal{N}(0,1)$.
- χ^2-distributed Q with k degrees of freedom $\iff$ $Q \sim \chi^2_k$:

$$Q = \sum_{i=1}^{k} Z_i^2.$$

- Mean:

$$\mathsf{E}(Q) = k.$$

- Variance:

$$\mathsf{V}(Q) = 2k.$$

- If $Q_1, \ldots, Q_m$ are independent and distributed as $\chi^2_{k_1}, \ldots, \chi^2_{k_m}$ respectively, then:

$$Q_1 + \ldots + Q_m \sim \chi^2_{k_1+\ldots+k_m}.$$

Critical values of χ^2_k-distribution

$k = 1, \ldots, 22$ (degrees of freedom)

DF	Right tail probability									
	0.995	0.99	0.975	0.95	0.9	0.5	0.1	0.05	0.025	0.01
1	0.0000	0.0002	0.0010	0.0039	0.0158	0.4549	2.7055	3.8415	5.0239	6.6349
2	0.0100	0.0201	0.0506	0.1026	0.2107	1.3863	4.6052	5.9915	7.3778	9.2103
3	0.0717	0.1148	0.2158	0.3518	0.5844	2.3660	6.2514	7.8147	9.3484	11.3449
4	0.2070	0.2971	0.4844	0.7107	1.0636	3.3567	7.7794	9.4877	11.1433	13.2767
5	0.4117	0.5543	0.8312	1.1455	1.6103	4.3515	9.2364	11.0705	12.8325	15.0863
6	0.6757	0.8721	1.2373	1.6354	2.2041	5.3481	10.6446	12.5916	14.4494	16.8119
7	0.9893	1.2390	1.6899	2.1673	2.8331	6.3458	12.0170	14.0671	16.0128	18.4753
8	1.3444	1.6465	2.1797	2.7326	3.4895	7.3441	13.3616	15.5073	17.5345	20.0902
9	1.7349	2.0879	2.7004	3.3251	4.1682	8.3428	14.6837	16.9190	19.0228	21.6660
10	2.1559	2.5582	3.2470	3.9403	4.8652	9.3418	15.9872	18.3070	20.4832	23.2093
11	2.6032	3.0535	3.8157	4.5748	5.5778	10.3410	17.2750	19.6751	21.9200	24.7250
12	3.0738	3.5706	4.4038	5.2260	6.3038	11.3403	18.5493	21.0261	23.3367	26.2170
13	3.5650	4.1069	5.0088	5.8919	7.0415	12.3398	19.8119	22.3620	24.7356	27.6882
14	4.0747	4.6604	5.6287	6.5706	7.7895	13.3393	21.0641	23.6848	26.1189	29.1412
15	4.6009	5.2293	6.2621	7.2609	8.5468	14.3389	22.3071	24.9958	27.4884	30.5779
16	5.1422	5.8122	6.9077	7.9616	9.3122	15.3385	23.5418	26.2962	28.8454	31.9999
17	5.6972	6.4078	7.5642	8.6718	10.0852	16.3382	24.7690	27.5871	30.1910	33.4087
18	6.2648	7.0149	8.2307	9.3905	10.8649	17.3379	25.9894	28.8693	31.5264	34.8053
19	6.8440	7.6327	8.9065	10.1170	11.6509	18.3377	27.2036	30.1435	32.8523	36.1909
20	7.4338	8.2604	9.5908	10.8508	12.4426	19.3374	28.4120	31.4104	34.1696	37.5662
22	8.6427	9.5425	10.9823	12.3380	14.0415	21.3370	30.8133	33.9244	36.7807	40.2894

Critical values of χ^2_k-distribution

$k = 24, \ldots, 100$ (degrees of freedom)

DF	Right tail probability									
	0.995	0.99	0.975	0.95	0.9	0.5	0.1	0.05	0.025	0.01
24	9.8862	10.8564	12.4012	13.8484	15.6587	23.3367	33.1962	36.4150	39.3641	42.9798
26	11.1602	12.1981	13.8439	15.3792	17.2919	25.3365	35.5632	38.8851	41.9232	45.6417
28	12.4613	13.5647	15.3079	16.9279	18.9392	27.3362	37.9159	41.3371	44.4608	48.2782
30	13.7867	14.9535	16.7908	18.4927	20.5992	29.3360	40.2560	43.7730	46.9792	50.8922
32	15.1340	16.3622	18.2908	20.0719	22.2706	31.3359	42.5847	46.1943	49.4804	53.4858
34	16.5013	17.7891	19.8063	21.6643	23.9523	33.3357	44.9032	48.6024	51.9660	56.0609
36	17.8867	19.2327	21.3359	23.2686	25.6433	35.3356	47.2122	50.9985	54.4373	58.6192
38	19.2889	20.6914	22.8785	24.8839	27.3430	37.3355	49.5126	53.3835	56.8955	61.1621
42	22.1385	23.6501	25.9987	28.1440	30.7654	41.3352	54.0902	58.1240	61.7768	66.2062
46	25.0413	26.6572	29.1601	31.4390	34.2152	45.3351	58.6405	62.8296	66.6165	71.2014
50	27.9907	29.7067	32.3574	34.7643	37.6886	49.3349	63.1671	67.5048	71.4202	76.1539
55	31.7348	33.5705	36.3981	38.9580	42.0596	54.3348	68.7962	73.3115	77.3805	82.2921
60	35.5345	37.4849	40.4817	43.1880	46.4589	59.3347	74.3970	79.0819	83.2977	88.3794
65	39.3831	41.4436	44.6030	47.4496	50.8829	64.3346	79.9730	84.8206	89.1771	94.4221
70	43.2752	45.4417	48.7576	51.7393	55.3289	69.3345	85.5270	90.5312	95.0232	100.4252
75	47.2060	49.4750	52.9419	56.0541	59.7946	74.3344	91.0615	96.2167	100.8393	106.3929
80	51.1719	53.5401	57.1532	60.3915	64.2778	79.3343	96.5782	101.8795	106.6286	112.3288
85	55.1696	57.6339	61.3888	64.7494	68.7772	84.3343	102.0789	107.5217	112.3934	118.2357
90	59.1963	61.7541	65.6466	69.1260	73.2911	89.3342	107.5650	113.1453	118.1359	124.1163
95	63.2496	65.8984	69.9249	73.5198	77.8184	94.3342	113.0377	118.7516	123.8580	129.9727
100	67.3276	70.0649	74.2219	77.9295	82.3581	99.3341	118.4980	124.3421	129.5612	135.8067

Problem 1

Let $X_1, \ldots, X_8$ be i.i.d. from $\mathcal{N}(0, 16)$, $Y_1, \ldots, Y_5$ be i.i.d. from $\mathcal{N}(1, 9)$, all X_i and Y_j are independent.

1. Find $\mathsf{P}\left(\sum_{i=1}^{8} X_i^2 < 20\right)$.
2. Find $\mathsf{P}\left(\sum_{j=1}^{5} (Y_j - 1)^2 > 12\right)$.
3. Could you find so easily $\mathsf{P}\left(\sum_{i=1}^{8} X_i^2 + \sum_{j=1}^{5} (Y_j - 1)^2 > 32\right)$?

Student's t-distribution

- Let $Z \sim \mathcal{N}(0,1)$ and $Q \sim \chi^2_k$ be independent random variables.
- t-distributed T with k degrees of freedom $\iff$ $T \sim t_k$:

$$T = \frac{Z}{\sqrt{Q/k}}.$$

- Mean:

$$\mathsf{E}(T) = 0, \quad k > 1.$$

- Variance:

$$\mathsf{V}(T) = \frac{k}{k-2}, \quad k > 2.$$

- Increasing degrees of freedom:

$$T \xrightarrow[k\to\infty]{d} Z.$$

Critical values of t_k-distribution

$k = 1, \ldots, 18$ (degrees of freedom)

DF	Right tail probability									
	0.25	0.20	0.15	0.10	0.05	0.025	0.01	0.005	0.001	0.0005
1	1.0000	1.3764	1.9626	3.0777	6.3138	12.7062	31.8205	63.6567	318.3088	636.6192
2	0.8165	1.0607	1.3862	1.8856	2.9200	4.3027	6.9646	9.9248	22.3271	31.5991
3	0.7649	0.9785	1.2498	1.6377	2.3534	3.1824	4.5407	5.8409	10.2145	12.9240
4	0.7407	0.9410	1.1896	1.5332	2.1318	2.7764	3.7469	4.6041	7.1732	8.6103
5	0.7267	0.9195	1.1558	1.4759	2.0150	2.5706	3.3649	4.0321	5.8934	6.8688
6	0.7176	0.9057	1.1342	1.4398	1.9432	2.4469	3.1427	3.7074	5.2076	5.9588
7	0.7111	0.8960	1.1192	1.4149	1.8946	2.3646	2.9980	3.4995	4.7853	5.4079
8	0.7064	0.8889	1.1081	1.3968	1.8595	2.3060	2.8965	3.3554	4.5008	5.0413
9	0.7027	0.8834	1.0997	1.3830	1.8331	2.2622	2.8214	3.2498	4.2968	4.7809
10	0.6998	0.8791	1.0931	1.3722	1.8125	2.2281	2.7638	3.1693	4.1437	4.5869
11	0.6974	0.8755	1.0877	1.3634	1.7959	2.2010	2.7181	3.1058	4.0247	4.4370
12	0.6955	0.8726	1.0832	1.3562	1.7823	2.1788	2.6810	3.0545	3.9296	4.3178
13	0.6938	0.8702	1.0795	1.3502	1.7709	2.1604	2.6503	3.0123	3.8520	4.2208
14	0.6924	0.8681	1.0763	1.3450	1.7613	2.1448	2.6245	2.9768	3.7874	4.1405
15	0.6912	0.8662	1.0735	1.3406	1.7531	2.1314	2.6025	2.9467	3.7328	4.0728
16	0.6901	0.8647	1.0711	1.3368	1.7459	2.1199	2.5835	2.9208	3.6862	4.0150
17	0.6892	0.8633	1.0690	1.3334	1.7396	2.1098	2.5669	2.8982	3.6458	3.9651
18	0.6884	0.8620	1.0672	1.3304	1.7341	2.1009	2.5524	2.8784	3.6105	3.9216

Critical values of t_k-distribution

$k = 19, \ldots, \infty$ (degrees of freedom)

DF	Right tail probability									
	0.25	0.20	0.15	0.10	0.05	0.025	0.01	0.005	0.001	0.0005
19	0.6876	0.8610	1.0655	1.3277	1.7291	2.0930	2.5395	2.8609	3.5794	3.8834
20	0.6870	0.8600	1.0640	1.3253	1.7247	2.0860	2.5280	2.8453	3.5518	3.8495
21	0.6864	0.8591	1.0627	1.3232	1.7207	2.0796	2.5176	2.8314	3.5272	3.8193
22	0.6858	0.8583	1.0614	1.3212	1.7171	2.0739	2.5083	2.8188	3.5050	3.7921
23	0.6853	0.8575	1.0603	1.3195	1.7139	2.0687	2.4999	2.8073	3.4850	3.7676
24	0.6848	0.8569	1.0593	1.3178	1.7109	2.0639	2.4922	2.7969	3.4668	3.7454
25	0.6844	0.8562	1.0584	1.3163	1.7081	2.0595	2.4851	2.7874	3.4502	3.7251
26	0.6840	0.8557	1.0575	1.3150	1.7056	2.0555	2.4786	2.7787	3.4350	3.7066
27	0.6837	0.8551	1.0567	1.3137	1.7033	2.0518	2.4727	2.7707	3.4210	3.6896
28	0.6834	0.8546	1.0560	1.3125	1.7011	2.0484	2.4671	2.7633	3.4082	3.6739
29	0.6830	0.8542	1.0553	1.3114	1.6991	2.0452	2.4620	2.7564	3.3962	3.6594
30	0.6828	0.8538	1.0547	1.3104	1.6973	2.0423	2.4573	2.7500	3.3852	3.6460
40	0.6807	0.8507	1.0500	1.3031	1.6839	2.0211	2.4233	2.7045	3.3069	3.5510
60	0.6786	0.8477	1.0455	1.2958	1.6706	2.0003	2.3901	2.6603	3.2317	3.4602
80	0.6776	0.8461	1.0432	1.2922	1.6641	1.9901	2.3739	2.6387	3.1953	3.4163
100	0.6770	0.8452	1.0418	1.2901	1.6602	1.9840	2.3642	2.6259	3.1737	3.3905
1000	0.6747	0.8420	1.0370	1.2824	1.6464	1.9623	2.3301	2.5808	3.0984	3.3003
∞	0.6745	0.8416	1.0364	1.2816	1.6449	1.9600	2.3263	2.5758	3.0902	3.2905

Problem 2

Let $X_1, \ldots, X_8$ be i.i.d. from $\mathcal{N}(0, 16)$, $Y_1, \ldots, Y_5$ be i.i.d. from $\mathcal{N}(1, 9)$, all X_i and Y_j are independent.

1. Find $\mathsf{P}\left(X_1 < \sqrt{\sum_{j=1}^{5}(Y_j - 1)^2}\right)$.
2. Find $\mathsf{P}\left(X_1 + 2X_2 < \sqrt{\sum_{j=1}^{4}(Y_j - 1)^2}\right)$.
3. Find $\mathsf{P}\left(Y_1 - 1 < \sqrt{\sum_{i=1}^{8} X_i^2}\right)$.

Fisher-Snedecor's F-distribution

- Let $Q \sim \chi^2_p$ and $R \sim \chi^2_k$ be independent random variables.
- F-distributed F with degrees of freedom p and k $\iff$ $F \sim F_{p,k}$:

$$F = \frac{Q/p}{R/k}.$$

By composition, inverse of F:

$$\frac{1}{F} \sim F_{k,p}.$$

- If $T \sim t_k$ then:

$$T^2 \sim F_{1,k}.$$

- Mean:

$$\mathsf{E}(F) = \frac{k}{k-2}, \quad k > 2.$$

- Variance:

$$\mathsf{V}(F) = \frac{2k^2(p+k-2)}{p(k-2)^2(k-4)}, \quad k > 4.$$

Problem 3

Let $X_1, \ldots, X_8$ be i.i.d. from $\mathcal{N}(0, 16)$, $Y_1, \ldots, Y_5$ be i.i.d. from $\mathcal{N}(1, 9)$, all X_i and Y_j are independent.

1. Find $\mathsf{P}\left(\sum_{i=1}^{8} X_i^2 < \sum_{j=1}^{5} (Y_j - 1)^2\right)$.
2. Find $\mathsf{P}\left(\sum_{i=1}^{4} X_i^2 < 7\sum_{j=1}^{3} (Y_j - 1)^2\right)$.
3. Find $\mathsf{P}\left(\sum_{i=1}^{3} X_i^2 < \sum_{i=4}^{8} X_i^2\right)$.
4. Find $\mathsf{P}\left(\sum_{i=1}^{7} X_i^2 < \sum_{i=3}^{8} X_i^2\right)$.

Quantiles of F-distribution

- There is a connection between quantiles of $F \sim F_{p,k}$ and $\frac{1}{F} \sim F_{k,p}$:

$$\begin{aligned}
\mathsf{P}(F < x_p) &= p, \\
\mathsf{P}\left(\frac{1}{x_p} < \frac{1}{F}\right) &= p, \\
\mathsf{P}\left(\frac{1}{F} > \frac{1}{x_p}\right) &= p, \\
\mathsf{P}\left(\frac{1}{F} < \frac{1}{x_p}\right) &= 1 - p.
\end{aligned}$$

- In terms of quantile function $Q(p)$:

$$Q_{p,k}(p) = \frac{1}{Q_{k,p}(1-p)}.$$

Types of random variables convergence

- Let $X_1, \ldots, X_n : \Omega \to \mathbb{R}$.

Definition

Sequence $X_1, \ldots, X_n$ converges in probability to X:

$$X_n \xrightarrow[n\to\infty]{\mathsf{P}} X,$$

$$\text{if } \forall \varepsilon > 0 : \qquad \lim_{n\to\infty} \mathsf{P}(|X_n - X| \geq \varepsilon) = 0.$$

Definition

Sequence $X_1, \ldots, X_n$ converges almost surely to X:

$$X_n \xrightarrow[n\to\infty]{\text{a.s.}} X,$$

$$\text{if } \omega \in \Omega : \qquad \mathsf{P}\left(\lim_{n\to\infty} X_n(\omega) = X(\omega)\right) = 1.$$

Law of large number

- Let $X_1, \ldots, X_n$ be identically distributed and pairwise uncorrelated with $\mathsf{E}\left(X_i^2\right) < \infty$.

Theorem (Weak LLN)

$$\frac{X_1 + \ldots + X_n}{n} \xrightarrow[n\to\infty]{\mathsf{P}} \mathsf{E}(X_i).$$

- Let $X_1, \ldots, X_n$ be i.i.d. with $\mathsf{E}\left|X_i\right| < \infty$.

Theorem (Strong LLN)

$$\frac{X_1 + \ldots + X_n}{n} \xrightarrow[n\to\infty]{\text{a.s.}} \mathsf{E}(X_i).$$

Rate of convergence for LLN and CLT

- Comparison of rates for sequence of i.i.d. $X_1, \ldots, X_n$ with $\mathsf{E}(X_i) = \mu$ and $\mathsf{V}(X_i) = \sigma^2$:

$$\text{LLN:} \qquad \frac{X_1 + \ldots + X_n - n\mu}{n} \xrightarrow[n\to\infty]{\text{a.s.}} 0,$$

$$\text{CLT:} \qquad \frac{X_1 + \ldots + X_n - n\mu}{\sigma\sqrt{n}} \xrightarrow[n\to\infty]{d} Z,$$

where $Z \sim \mathcal{N}(0, 1)$.

Sample mean

- Let sample $X_1, \ldots, X_n$ from the same population be i.i.d. random variables with $\mathsf{E}(X_i) = \mu$ and $\mathsf{V}(X_i) = \sigma^2$.
- Sample mean $\overline{X}$:

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

- According to LLN, sample mean is a natural estimator of true population mean:

$$\mathsf{E}\left(\overline{X}\right) = \mu.$$

- Variance of sample mean decreases with the growth of sample:

$$\mathsf{V}\left(\overline{X}\right) = \frac{\sigma^2}{n}.$$

- If population was normal with $X_i \sim \mathcal{N}\left(\mu, \sigma^2\right)$:

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

Problem 4

Suppose that we plan to take a random sample of size n from a normal distribution with mean μ and standard deviation $\sigma = 2$.

1. Suppose $\mu = 4$ and $n = 20$.
   1. What is the probability that the mean $\overline{X}$ of the sample is greater than 5?
   2. What is the probability that $\overline{X}$ is smaller than 3?
   3. What is $\mathsf{P}\left(\left|\overline{X} - \mu\right| \leq 1\right)$ in this case?
2. How large should n be in order that $\mathsf{P}\left(\left|\overline{X} - \mu\right| \leq 0.5\right) \geq 0.95$ for every possible value of μ?
3. It is claimed that the true value of μ is 5 in a population. A random sample of size $n = 100$ is collected from this population, and the mean for this sample is $\overline{x} = 5.8$. Based on the result in (2), what would you conclude from this value of $\overline{X}$?

CLT for sample mean

- In case when population X_i is NOT normal, sample distribution of $\overline{X}$ still can be approximated with normal distribution:

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

  for sample size $n > 30$.
- This is due to natural application of CLT:

$$\frac{X_1 + \ldots + X_n - n\mu}{\sigma\sqrt{n}} \xrightarrow[n\to\infty]{d} Z,$$

$$\frac{\frac{X_1 + \ldots + X_n}{n} - \mu}{\sigma/\sqrt{n}} \xrightarrow[n\to\infty]{d} Z,$$

$$\frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \xrightarrow[n\to\infty]{d} Z.$$

Problem 5

Suppose $X_1, X_2, \ldots, X_{40}$ are i.i.d. random variables with c.d.f.

$$F(x) = \begin{cases} 0, & x < 0, \\ x^3, & 0 \le x \le 1, \\ 1, & x > 1. \end{cases}$$

1. Find $\mathsf{P}\left(X_1 > 0.8 \mid X_1 > 0.5\right)$ and $\mathsf{P}\left(X_1 > 0.8 \mid X_2 > 0.5\right)$.
2. Let the mean be $\overline{X} = \dfrac{1}{40} \sum\limits_{i=1}^{40} X_i$. Estimate probability $\mathsf{P}\left(\overline{X} > 0.7\right)$.

Look at the time!