REGIONAL TRAINING COURSE 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
5 June 2018

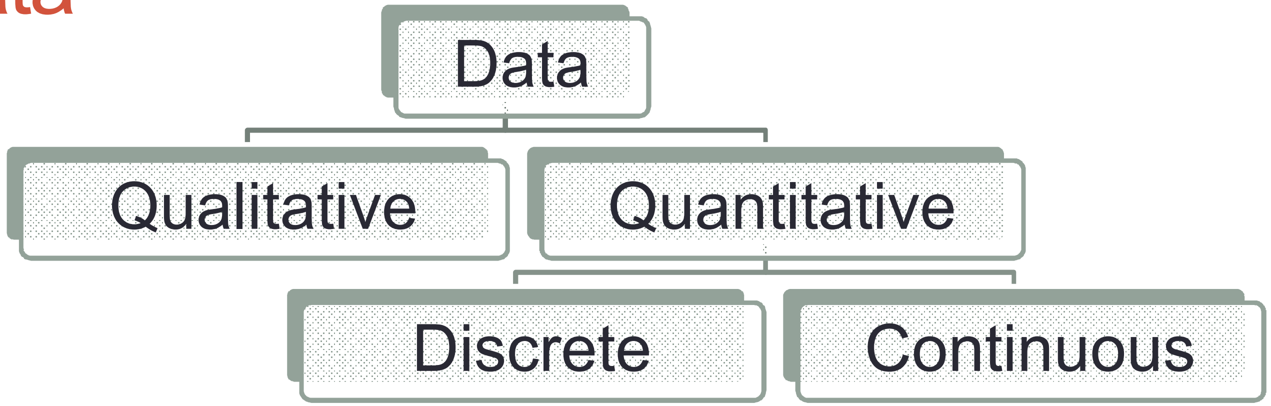
Outline

- **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 2 Outline

- Random variable (discrete and continuous)
- Probability density function (PDF)
- Probability distribution (discrete and continuous)
- Expected value, Variance, Covariance, Correlation
- Discrete probability distribution
  - Binomial probability distribution
  - Poisson probability distribution

Types of data



- Qualitative data provide names of items
- Quantitative data provide numerical description of items
- Discrete data can take only integer (whole) values (result of counting)
- Continuous data can take any values in a certain interval (decimal is possible)

Examples:

1) Discrete: a number of products (customers, firms, etc): 5, 8, 13, etc
2) Continuous: the weight of products: 5kg, 5.3kg, 5.235kg, etc.

Exercise: Classify into discrete or continuous:

Number of children, time, height, sales*, votes*, expenditure.

Random variable

- A numerical description of an outcome of an experiment
- Can be discrete or continuous
- Answers the question “how many?” or “how much?”

Examples:

1) A store has 20 tables. The random variable X is a number of tables sold. Then X = {0,1,2,…,19,20}
2) A store has 20 kg of sugar. The random variable X is the amount of sugar sold. Then $0 \leq X \leq 20$.

Exercise:

1) A sample of 30 out of 200 clients were questioned to choose between plan A or B. X defines the number of sampled clients who chose A. Find the range of X.
2) A customer has $1000 to spend at a grocery store. The random variable X defines the customer’s expenditure. Find the range of X.

Probability function and probability distributions

The **probability density function (PDF)**, denoted by $f(x)$, provides the probability of occurrence of a random variable (r. v.).

A **probability distribution** is a table, graph, or mathematical formula that shows all possible values of the random variable x and the associated probability function $f(x)$.

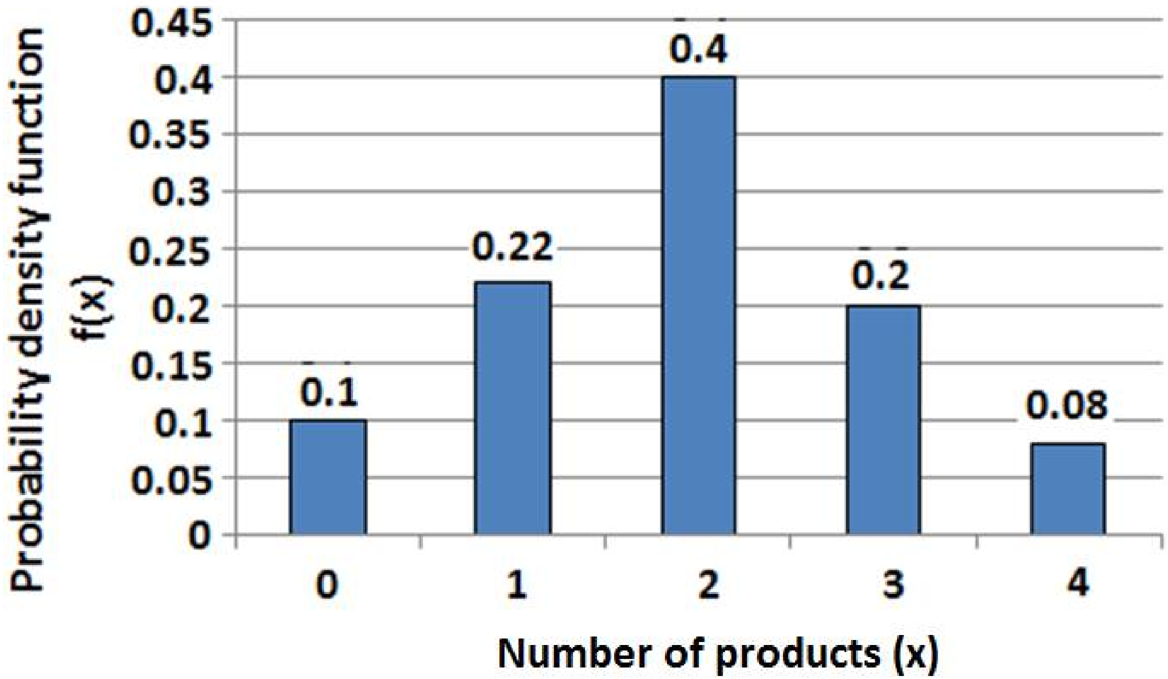
Two types of a probability distribution:

1) Discrete probability distribution (when r. v. is discrete)
2) Continuous probability distribution (when r. v. is continuous)

Example 1 (Discrete probability distribution)

Number of products purchased by customers:

# of prod. x	# of cust. f	Rel. freq. $f / \sum f$	f (x)	P(X=x)
0	10	0.10	0.10	0.10
1	22	0.22	0.22	0.22
2	40	0.40	0.40	0.40
3	20	0.20	0.20	0.20
4	8	0.08	0.08	0.08
	100	**1**	**1**	**1**



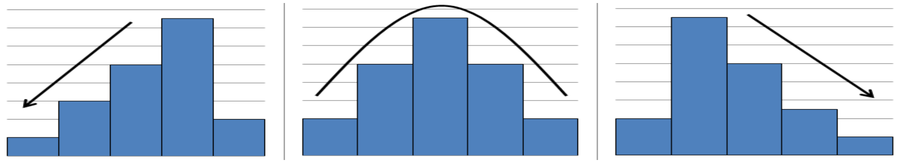
Conditions: $f(x) \geq 0, \sum_{i=1}^{n} f(x_i) = 1.$

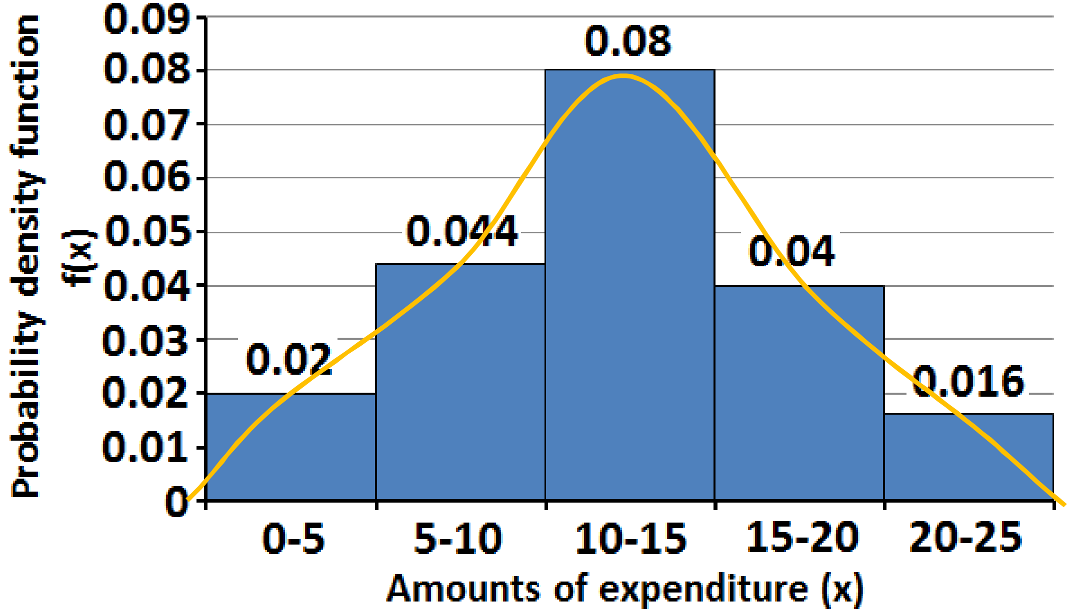
Exercise: Find **1)** P(X=2) = f (2), **2)** P(1≤X≤3), **3)** P(X<4)

Example (Continuous probability distribution)

Amounts of expenditure by customers:

x	# of customers	Relative frequency $f / \sum f$	f (x)
0-5	10	0.10	0.020
5-10	22	0.22	0.044
10-15	40	0.40	0.080
15-20	20	0.20	0.040
20-25	8	0.08	0.016
	100	**1**	





Conditions: $f(x) \geq 0,\ \int_{-\infty}^{+\infty} f(x)\,dx = 1.$

Exercise: Find **1)** P(15≤X<20), **2)** P(X=7), **3)** P(8≤X<12)

Expected value and Variance (Standard deviation)

Number of products purchased by customers

x	f	$P(X=x)$	$f(x)$	fx	xP	fx^2	$f(x-\bar{x})^2$	x^2P
0	10	0.10	0.10	0	0	0	37.64	0
1	22	0.22	0.22	22	0.22	22	19.44	0.22
2	40	0.40	0.40	80	0.80	160	0.14	1.60
3	20	0.20	0.20	60	0.60	180	22.47	1.80
4	8	0.08	0.08	32	0.32	128	33.95	1.28
	100	**1**	**1**	**194**	**1.94**	**490**	**113.64**	**4.90**

Statistics:

Mean = $\mu = \frac{\sum fx}{\sum f} = 1.94$

Var(X) = $\sigma^2 = \frac{\sum f\cdot(x-\bar{x})^2}{\sum f} = 1.1364$

Var(X) = $\frac{\sum fx^2}{\sum f} - \left(\frac{\sum fx}{\sum f}\right)^2 =$

$\frac{490}{100} - 1.94^2 = 1.1364$

SD = $\sigma = \sqrt{\sigma^2} = 1.07$

Probability:

E(X) = $\sum xP = 1.94$

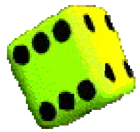
Var(X) = $E\left(\left(X - E(X)\right)^2\right) = E(X^2) - \left(E(X)\right)^2 = \sum x^2P - \left(\sum xP\right)^2 = 1.1364$

Exercise: Prove a) $\frac{\sum fx}{\sum f} = \sum xP$; b) $E\left(\left(X - E(X)\right)^2\right) = E(X^2) - \left(E(X)\right)^2$.

Exercise (Expected value and Variance)

In the game of rolling a die, a player wins \$10 if he gets 1, \$20 if 2, \$30 if 3, \$40 if 4, \$50 if 5, but he loses \$120 if he gets 6.

Find the expected value and variance of your net gain per play.



x	$P(X=x)$	xP	x^2P
10	1/6	10/6	100/6
20	1/6	20/6	400/6
30	1/6	30/6	900/6
40	1/6	40/6	1600/6
50	1/6	50/6	2500/6
-120	1/6	-20	2400
		5	**3316.67**

$$E(X) = \mu = \sum X \cdot P(X = x) = 5$$

$$V(X) = \sigma^2 = E(X^2) - (E(X))^2 = \sum x^2 P - \left(\sum xP\right)^2 = 3316.67 - 5^2 = 3291.67$$

$$SD = \sigma = 57.37$$

Covariance

Number of ads (x) and sales (y) in 5 stores

	x	y	$(x-\mu_x)(y-\mu_y)$	xy	x^2	y^2
1	1	22	5.4	22	1	484
2	1	24	1.8	24	1	576
3	4	25	0	100	16	625
4	5	28	6.6	140	25	784
5	3	26	0.2	78	9	676
Σ	**14**	**125**	**14**	**364**	**52**	**3145**
μ	**2.8**	**25**		**72.8**		
σ	**1.6**	**2**				

Statistics:

Cov (X,Y) $= \sigma_{xy} = \frac{\sum(x-\mu_x)(y-\mu_y)}{N} = \frac{14}{5} = 2.8$

Cov (X,Y) $= \sigma_{xy} = \frac{\sum xy - \frac{\sum x \sum y}{N}}{N} = \frac{14}{5} = 2.8$

Correl (X,Y) $= \frac{\sigma_{xy}}{\sigma_x \sigma_y} = \frac{\sum xy - \frac{\sum x \sum y}{N}}{\sqrt{\left(\sum x^2 - \frac{(\sum x)^2}{N}\right)\left(\sum y^2 - \frac{(\sum y)^2}{N}\right)}} =$

$$= \frac{14}{\sqrt{\left(52 - \frac{14^2}{5}\right)\left(3145 - \frac{125^2}{5}\right)}} = 0.875$$

Probability:

Cov (X,Y) $= \sigma_{xy} = E\left(\left(X - E(X)\right)\left(Y - E(Y)\right)\right) = E(XY) - E(X)E(Y) = 72.8 - 2.8 \cdot 25 = 2.8$

Covariance (continued)

Number of ads (x) and sales (y) in 5 stores

$x \setminus y$	22	24	25	28	26	$\mu_y = \mathbf{25}$
1	$\frac{1}{5}$	0	0	0	0	
1	0	$\frac{1}{5}$	0	0	0	
4	0	0	$\frac{1}{5}$	0	0	
5	0	0	0	$\frac{1}{5}$	0	
3	0	0	0	0	$\frac{1}{5}$	
$\mu_x = \mathbf{2.8}$						

Probability:

Cov (X,Y) $= \sum(x-\mu_x)(y-\mu_y)h(x,y) =$

$$= \sum xyh(x,y) - \mu_X\mu_Y =$$

$$= 1\cdot 22\cdot\frac{1}{5} + 1\cdot 24\cdot\frac{1}{5} + 4\cdot 25\cdot\frac{1}{5} +$$

$$= 5\cdot 28\cdot\frac{1}{5} + 3\cdot 26\cdot\frac{1}{5} - 2.8\cdot 25 = 2.8$$

Homework: Prove:

$$\sum(x-\mu_X)(y-\mu_Y)h(x,y) = \sum xyh(x,y) - \mu_X\mu_Y$$

Types of probability distributions

Two types of a probability distribution:

1) Discrete probability distribution:
   - Binomial
   - Poisson
   - Hypergeometric

2) Continuous probability distribution:
   - Uniform
   - Normal
   - F
   - Chi-squared

Binomial experiment and Binomial distribution

❖ Shows the number of occurrences in a multiple step experiment

Binomial experiment has 4 assumptions:

1) It has a sequence of n trials
2) Two outcomes are possible:
   Success (p) and Failure (1 – p)
3) The probability of a success does not change from trial to trial.
4) Trials are independent

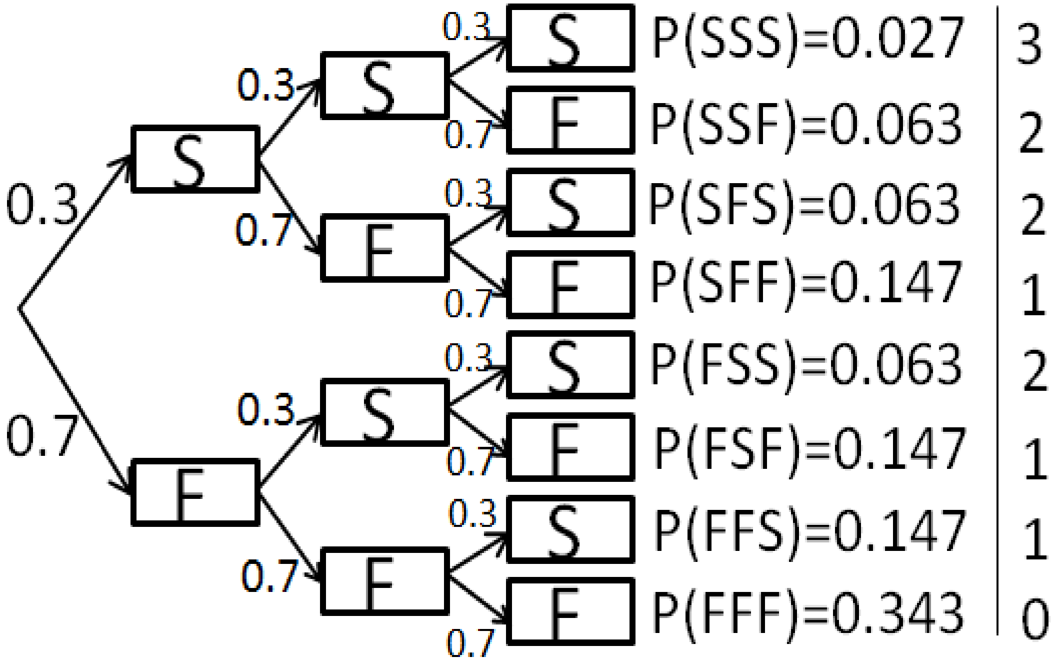
Example: Assume 3 customers enter the store during an hour and the probability of customer purchase is 0.3. Find:

a) $P(X = 1)$; b) $P(X \leq 1)$; c) $P(X \geq 1)$; d) $E(X)$; e) $Var(X)$.

Binomial experiment (distribution) I

Assume 3 customers enter the store during an hour and the probability of customer purchase is 0.3. Find:

a) $P(X=1)$; b) $P(X \le 1)$; c) $P(X \ge 1)$; d) $E(X)$; e) $Var(X)$.



x	**f (x)**
0	0.343
1	0.441
2	0.189
3	0.027
	1

a) $P(X=1) =$ $f(1) = 0.441;$

b) $P(X \le 1) =$ $f(0)+f(1) = 0.784;$

c) $P(X \ge 1) =$ $f(1)+f(2)+f(3) =$ $1 - f(0) = 0.657.$

d) $E(X) = 3 \cdot 0.3 = 0.9;$

e) $Var(X) = 3 \cdot 0.3 \cdot 0.7 = 0.63.$

$$E(X) = \sum xp = np$$

$$Var(X) = np(1-p)$$

Binomial experiment (distribution) II

1-method: $f(x) = C_x^n p^x (1-p)^{n-x}$

$$f(2) = C_2^3 p^2 (1-p)^{3-2} = \frac{3!}{2!\,(3-2)!} 0.3^2 (1-0.3) = 0.189$$

Exercise: Calculate $f(0)$.

2-method: Binomial Table (see next slide)

n = 3, x = 2, p = 0.3. Hence: $f(2) = P(X=2) = 0.189$

3-method: In MS Excel: BINOM.DIST(x,n,p,0)

BINOM.DIST(2,3,0.3,0) = 0.189

x	**f (x)**
0	0.343
1	0.441
2	0.189
3	0.027
	1

Binomial probability table

								p								
n	***x***	**0.25**	**0.30**	**0.35**	**0.40**	**0.45**	**0.50**	**0.55**	**0.60**	**0.65**	**0.70**	**0.75**	**0.80**	**0.85**	**0.90**	**0.95**
2	**0**	.5625	.4900	.4225	.3600	.3025	.2500	.2025	.1600	.1225	.0900	.0625	.0400	.0225	.0100	.0025
	1	.3750	.4200	.4550	.4800	.4950	.5000	.4950	.4800	.4550	.4200	.3750	.3200	.2550	.1800	.0950
	2	.0625	.0900	.1225	.1600	.2025	.2500	.3025	.3600	.4225	.4900	.5625	.6400	.7225	.8100	.9025
3	**0**	.4219	.3430	.2746	.2160	.1664	.1250	.0911	.0640	.0429	.0270	.0156	.0080	.0034	.0010	.0001
	1	.4219	.4410	.4436	.4320	.4084	.3750	.3341	.2880	.2389	.1890	.1406	.0960	.0574	.0270	.0071
	2	.1406	**.1890**	.2389	.2880	.3341	.3750	.4084	.4320	.4436	.4410	.4219	.3840	.3251	.2430	.1354
	3	.0156	.0270	.0429	.0640	.0911	.1250	.1664	.2160	.2746	.3430	.4219	.5120	.6141	.7290	.8574
4	**0**	.3164	.2401	.1785	.1296	.0915	.0625	.0410	.0256	.0150	.0081	.0039	.0016	.0005	.0001	.0000
	1	.4219	.4116	.3845	.3456	.2995	.2500	.2005	.1536	.1115	.0756	.0469	.0256	.0115	.0036	.0005
	2	.2109	.2646	.3105	.3456	.3675	.3750	.3675	.3456	.3105	.2646	.2109	.1536	.0975	.0486	.0135
	3	.0469	.0756	.1115	.1536	.2005	.2500	.2995	.3456	.3845	.4116	.4219	.4096	.3685	.2916	.1715
	4	.0039	.0081	.0150	.0256	.0410	.0625	.0915	.1296	.1785	.2401	.3164	.4096	.5220	.6561	.8185
5	**0**	.2373	.1681	.1160	.0778	.0503	.0312	.0185	.0102	.0053	.0024	.0010	.0003	.0001	.0000	.0000

Poisson experiment and Binomial distribution

❖ Shows the number of occurrences over a certain time or space

Poisson experiment has 2 assumptions:

1) The probability of an occurrence is the same for any two intervals of equal length
2) Occurrences in the intervals are independent

Example:

Phone calls arrive at the rate of 4 calls per 5 minutes at the reservation desk of UzAir company. Find:

a) $P(X = 3)$; b) $P(X = 0)$; c) $P(X \leq 1)$.

Poisson experiment (distribution) I

Phone calls arrive at the rate of 4 calls per 5 minutes at the reservation desk of UzAir company. Find:

a) $P(X = 3)$; b) $P(X = 0)$; c) $P(X \leq 1)$.

This is a Poisson experiment, because:

1) The probability of phone calls (4) is the same for any two periods of equal length
2) Occurrence or non-occurrence of a phone call in any period is independent of the occurrence or non-occurrence in any other period.

Poisson experiment (distribution) II

1-method: $f(x) = \frac{\mu^x e^{-\mu}}{x!}$, where $e = 2.718 \ldots$

$$f(3) = \frac{4^3 e^{-4}}{3!} = \frac{64 \cdot 0.0183}{6} = 0.1954$$

Exercise: Calculate: b) $f(0)$; c) $P(X \leq 1)$.

2-method: Poisson Table (see next slide)

$\mu = 4,\ x = 3, f(3) = P(X = 3) = 0.1954$

3-method: In MS Excel: POISSON.DIST(x,μ,0)

POISSON.DIST(3,4,0) = 0.1954.

Poisson probability table

	μ														
x	**2.1**	**2.2**	**2.3**	**2.4**	**2.5**	.	**3.9**	**4.0**	**4.1**	.	**11**	**12**	**13**	**14**	**15**
0	.1225	.1108	.1003	.0907	.0821	.	.0202	.0183	.0166	.	.0000	.0000	.0000	.0000	.0000
1	.2572	.2438	.2306	.2177	.2052	.	.0789	.0733	.0679	.	.0002	.0001	.0000	.0000	.0000
2	.2700	.2681	.2652	.2613	.2565	.	.1539	.1465	.1393	.	.0010	.0004	.0002	.0001	.0000
3	.1890	.1966	.2033	.2090	.2138	.	.2001	**.1954**	.1904	.	.0037	.0018	.0008	.0004	.0002
4	.0992	.1082	.1169	.1254	.1336	.	.1951	.1954	.1951	.	.0102	.0053	.0027	.0013	.0006
5	.0417	.0476	.0538	.0602	.0668	.	.1522	.1563	.1600	.	.0224	.0127	.0070	.0037	.0019
6	.0146	.0174	.0206	.0241	.0278	.	.0989	.1042	.1093	.	.0411	.0255	.0152	.0087	.0048
7	.0044	.0055	.0068	.0083	.0099	.	.0551	.0595	.0640	.	.0646	.0437	.0281	.0174	.0104
8	.0011	.0015	.0019	.0025	.0031	.	.0269	.0298	.0328	.	.0888	.0655	.0457	.0304	.0194
9	.0003	.0004	.0005	.0007	.0009	.	.0116	.0132	.0150	.	.1085	.0874	.0661	.0473	.0324
10	.0001	.0001	.0001	.0002	.0002	.	.0045	.0053	.0061	.	.1194	.1048	.0859	.0663	.0486

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.