Examination Number:____________

Sign the Honor Pledge Below_____________
____

Last Name______________First__________
PID #_________________________________
Write Your Section Number here:_________

University of North Carolina
Economics 400: Economic Statistics & Econometrics
First Midterm Examination

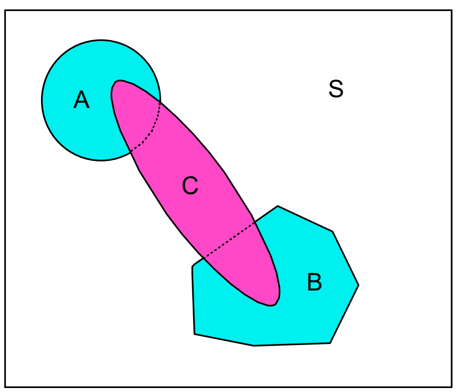
Prof. B. Turchi

General Instructions: Answer all 9 questions on this examination, writing your answers on the exam paper itself. Use the back of the pages for any extra work, if necessary. Sign the Honor Pledge above. Express all answers to a precision of at least 3 decimal points. Show your work to be eligible for partial credit. **Be sure to note that tables and formulas are on the last 2 pages of the exam.**

Part I: Each question in this part (6 questions in all) is worth 5 points.

1. (5 points) Suppose we have three events *A*, *B,* and *C* in a sample space, *S* as shown in the Venn diagram below. Write the expression for the probability of event C in terms of the sample space and events *A* and *B*. *P(C)* = __________________.

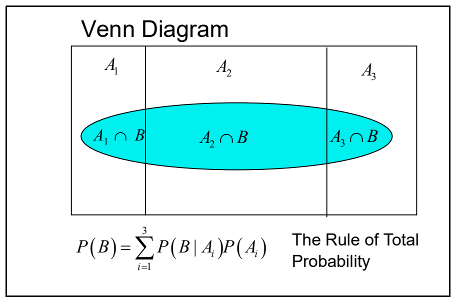
$P(C) = P(C \cap A) + P(C \cap \bar{A})$ or,
$P(C) == P(C \cap B) + P(C \cap \bar{B})$



There may be other ways to calculate P(C), so be sure to check if student has used an alternate but correct method

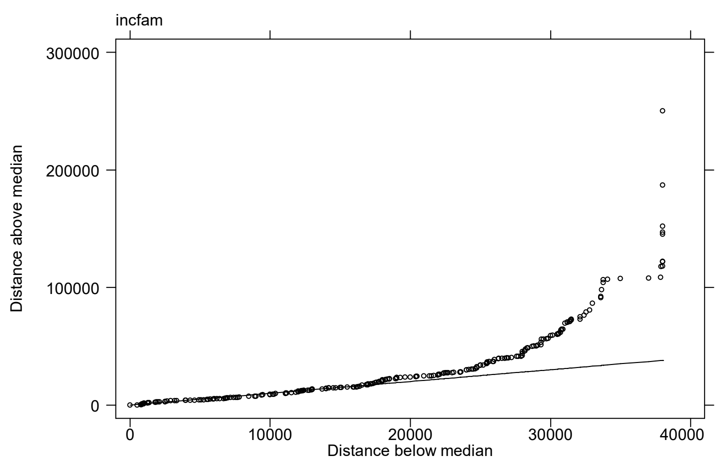
2. (5 points) Suppose we have events A_1, A_2, A_3, that are exhaustive and mutually exclusive, then for any event, B, complete the expression, and draw a Venn diagram illustrating this situation.

$$P(B) = \sum_{i=1}^{3} P(B \mid A_i) P(A_i)$$

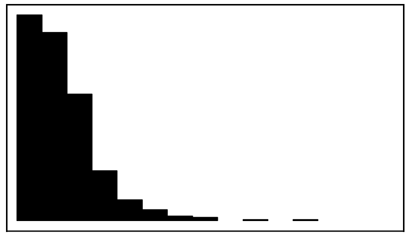


The student may not have extended set B across all 3 values of A; however that only means that one of the conditional probabilities would be zero.

3. (5 points) Often we collect data and want to get an idea of the shape of the empirical distribution of those data. Below we have a **symmetry plot** for family income data drawn from a sample survey of households.



Describe the shape of this sample of family income data. Is it symmetrical? skewed (if so, in which direction)? Sketch a histogram of these data in the box to illustrate your answer



These data are highly skewed to the right.

4. (5 points) Which of the following Stata commands prints out on the log the highest and lowest values of the variable *score* in a data set containing many variables? *(Circle correct answer)*

```
list score in 30/80
list if (score <30 | score >80)
list if (score >30 | score >80)
list if (score <30 & score > 80)
```

5. (5 points) Write below the Stata command that would give you the descriptive statistics for the variable *score,* including the mean, the median, the quartiles and the extreme values.

summarize score, detail

6. (5 points) In how many states does annual beer consumption exceed 22.9 gallons per capita?

23 states

```
. stem galperca

Stem-and-leaf plot for galperca (State Beer Consumption- Gal. Per Capita)

galperca rounded to nearest multiple of .1
plot in units of .1

  1** | 30
  1** |
  1** | 73,75,79
  1** | 88,93,93,95,98
  2** | 00,01,01,03,05,07,07,08,10,10,13,13,16,18
  2** | 21,22,25,29,29,30,33,35,35
  2** | 40,41,44,44,46,49,49,52,56
  2** | 64,67,69,70,70,70,78
  2** | 86
  3** | 13
  3** |
  3** | 44
```

Examination Number:____________

Part II: The next 3 questions have multiple parts and are worth 70 points in total; be sure to answer all parts.

7. (25 points total) Despite the national furor over the arrival of the West Nile virus, the number of severe reported cases (projected to be 3,619 in 2002) is relatively small and the number of deaths resulting from those cases is also projected to be small (192 deaths). The West Nile virus is transmitted to humans through the bite of an infected mosquito. At present only about 0.5 percent of all mosquitos are infected with West Nile virus. Use the following information to answer the questions below:

Probability of being bitten by a mosquito: 0.35
Probability of catching the disease if bitten: 0.005
Probability of having a severe case if infected: 0.002
Probability of dying if have a severe case: 0.09

Let:

B = bitten by a mosquito
C = catch the disease
S = have a severe case
D = die from West Nile fever

a) (3 points) Write symbolically and numerically the probability of being bitten:

$\boxed{P(B) = 0.35}$

b) (3 points) Write symbolically and numerically the probability of catching the disease if bitten: $\boxed{P(C \mid B) = 0.005}$

c) (3 points) Write symbolically and numerically the probability of having a severe case if infected: $\boxed{P(S \mid C) = 0.002}$

d) (3 points) Write symbolically and numerically the probability of dying if have a severe case: $\boxed{P(D \mid S) = 0.09}$

e) (13 points) compute the probability [P(D)] that a randomly selected person will die from the West Nile virus in 2002. Show the requested probability statement in symbolic terms and then use the information above to compute the probability of dying. Be sure to show your work.

$$P(D)=P(D\mid S)P(S)+P(D\mid \bar{S})P(\bar{S})=P(D\mid S)P(S)+0$$

$$P(S)=P(S\mid C)P(C)$$

$$P(C)=P(C\mid B)P(B)$$

$$\Rightarrow P(D)=P(D\mid S)P(S\mid C)P(C\mid B)P(B)=0.005\times 0.002\times 0.09\times 0.35=3.15\times 10^{-7}$$

There are 2 ways to die from West Nile virus: (1) having been severely infected, and (2) not having been severely infected. $P(D|\bar{S})=0$ so the only way you can die is to have been severely infected first. The only way you can be severely infected is to have caught the disease in the first place, and the only way you can catch the disease is to have been bitten. Consequently, $P(D)$ is the product of three conditional probabilities, all of which we know, and the total probability of being bitten, which we also know (last equation in box above.)

Give credit for correct numerical answer, but don't give complete credit unless the student has given evidence that he/she understands the process of developing $P(D)$.

8. (25 points total) School districts pride themselves on the performance of their students on standardized tests of achievement and ability. A good example is the Advanced Placement Test for calculus. Suppose that the Wake County School district has 540 students enrolled in AP calculus. Based on prior years' experience we can expect 151 of these students to score a 4 or better on the AP calculus test. Suppose that we randomly sample 26 *different* students from the current group.

Here's the information available from the stem of the question:

r = 151

N = 540

n = 26

r/N = p = 0.28

We are sampling without replacement, so theoretically one should use the hypergeometric distribution. However, the rule of thumb n is less than 5 percent of N applies here, so we could conceivably use the binomial distribution to approximate the hypergeometric. Moreover, because the rules of thumb for approximating the binomial with the normal are also satisfied (np >= 5, n(1-p) >= 5) we can use the normal distribution to approximate the binomial. This really comes in handy in part c) of this question.

a) (4 points) What is the expected number of students in this sample who receive a 4 or better on the AP test?

Use hypergeometric to compute $E\left[\tilde{x}\right]=n\left(\frac{r}{N}\right)=26\left(\frac{151}{540}\right)=7.27$

Using the rule of thumb, some students might propose to use the binomial with p = 151/540 = 0.2796 and *np = 7.27.* This is acceptable, too.

b) (4 points) What is the variance of this sample?

Using the hpergeometric formula for variance: Variance: $\sigma^2=n\left(\frac{r}{N}\right)\left(\frac{N-r}{N}\right)\left(\frac{N-n}{N-1}\right)$

$$26(\tfrac{151}{540})(\tfrac{540-151}{540})(\tfrac{540-26}{539})=4.994.$$

Using the binomial: Variance: $\sigma^2=np(1-p)=26\times 0.2796\times 0.7204=5.237$ This approximation of the variance is a bit further off as compared to the expected values above. Go ahead and give full credit for the binomial version, too.

c) (9 points) What is the probability that between 6 and 13 students inclusive score a 4 or better on the exam?

Here we can take advantage of the fact that this problem meets the rule-of-thumb criteria for applying the normal distribution to a binomial distribution. So, using mean and variance from *either* the hypergeometric or binomial distributions as calculated above, calculate *z-scores* for random variable (x) values 5.5 and 13.5 (note, we've applied the continuity correction *before* computing z-scores):

Binomial:

$$z_{5.5}=\frac{5.5-7.27}{2.29}=-0.773\Rightarrow P\left(z_0-z_{0.773}\right)=0.28027$$
$$z_{13.5}=\frac{13.5-7.27}{2.29}=2.720\Rightarrow P\left(z_0-z_{2.720}\right)=0.4967$$
$$P\left(z_{-0.773}-z_{2.720}\right)=0.2803+0.4967=0.777$$

Hypergeometric:

$$z_{5.5}=\frac{5.5-7.27}{2.235}=-0.792\Rightarrow P\left(z_0-z_{0.792}\right)=0.2858$$
$$z_{13.5}=\frac{13.5-7.27}{2.235}=2.787\Rightarrow P\left(z_0-z_{2.787}\right)=0.4974$$
$$P\left(z_{-0.792}-z_{2.787}\right)=0.2858+0.4974=0.7832$$

Note that the student <u>could</u> have computed these probabilities using the hypergeometric or the binomial distributions; however, the work would have been extraordinarily tedious. If they calculate the probability <u>correctly</u> using either of these two discrete distributions give full credit.

d) (4 points) What is the theoretically most appropriate distribution to use in calculating the answer to part c)? Why?

Theoretically one might have used the hypergeometric distribution because of the sampling without replacement.

e) (4 points) What distribution did you use to calculate the answer to part c)? Why?

Here, the important point is to recognize that rules of thumb would have allowed approximation with the binomial, or, preferably, the normal distribution. If the student used the normal give full credit. If binomial, subtract one point, if hypergeometric subtract 2 points, if other method, subtract 3 points.

9. (20 points total) A couple of years ago, my children bought me a DVD player for Christmas. It had a one-year warranty and failed completely 376 days after purchase. In buying a new one, I was faced with the problem of buying an extended warranty that would extend warranty coverage through the 3d year of ownership (i.e., the extended warranty would expire on the 4th anniversary of purchase). The machine cost me $250 new and would (we will assume) cost me the same to replace if it should fail. Finally, the salesman at BestBuy (an economics student at UNC) told me that he had seen company data suggesting that the mean time before failure of these machines is 3.2 years.

a) (2 points) What is the probability that the machine will fail in *less than* zero years after purchase?

The probability is 0. The point of this question was to force the student to see that the random variable *time until failure* can only take on positive values. Hence, the only continuous probability distribution available for use is the *exponential distribution.*

b) (6 points) What is the probability that the machine will fail between 1 and 4 years?

This is a problem that requires use of the exponential distribution and, in particular, the formula for *right-tail probabilities:* $P(x \geq a) = e^{-\lambda a}$. We know that the mean of this distribution is $3.2 = \frac{1}{\lambda}$, so we have lambda. Use a= 1 and 4 and compute the two right tail probabilities, subtracting $P(x \geq 1) - P(x \geq 4) = 0.7316 - 0.2865 = 0.4451.$

c) (6 points) What is the *maximum* that I should pay for the extended warranty?

I should pay at a maximum the expected value of my loss: $250 \times 0.4451 = \$111.28$

d) (6 points) If BestBuy has to replace my machine during warranty, it will cost them $150 to do so. What is the *minimum* they will be willing to charge for the extended warranty?

BestBuy should be willing to charge, at a minimum, the expected value of their loss:

$150 \times 0.4451 = \$66.77$

10. (30 points) The owner of a local Mercedes Benz dealership is trying to sell the dealership to a buyer who wants to know the profit he can make if he buys the dealership. The dealer claims that he sells 1,200 cars per year at an average profit of $4,400 per car. The number of cars sold each year is easily established to be, in fact, 1,200; however, the profit per car is more difficult to establish and requires the buyer to hire an accountant to compute the actual profits on cars from a random sample of last year's sales. The buyer needs to be confident at the 95 percent level that the profit per car is at least $3,900 per car.

a) How many cars need to be sampled if we do not know whether the underlying distribution is normal, but we do know that the underlying standard deviation, σ, is $800 and we want a margin of error between the sample mean and population mean of no more than $200. What distribution will you assume for the sampling distribution of the sample mean? Why?

b) Assume that a random sample of the size you just computed in part a) is drawn and the sample mean profit is $4,110. Assuming the same population standard deviation as above, can we be 95 percent sure that the *minimum* profit is greater than or equal to $3,900? Show your work and sketch a diagram illustrating your calculation.

c) Is the dealer's claimed profit within the 95 percent confidence interval suggested by the above sample? (Again, sketch a diagram illustrating your answer)

Answers to Problem 10:

(a) This part of the question requires us to find a sample size that is sufficient to check two values: the claimed profit of $4,400 and the required profit of $3,900. However, some students may have used the lower one-sided confidence interval and focussed only on the $3,900 minimum profit requirement. This gets them into problems in parts b) and c).

Doing a standard 95% $(1-\alpha)$confidence interval we use the sample size formula:

$$n = z_{\frac{\alpha}{2}}^2 \frac{\sigma^2}{D^2} = 1.96^2 \frac{800^2}{200^2} = 61.47 \sim 62$$

Using the lower one-sided confidence interval we would have:

$$n = z_{\alpha}^{2}\frac{\sigma^{2}}{D^{2}} = 1.645^{2}\frac{800^{2}}{200^{2}} = 43.3 \sim 44$$

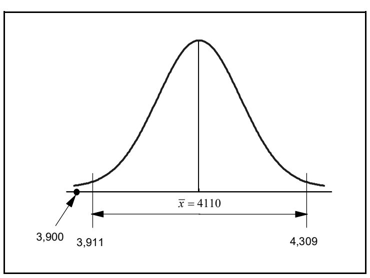
In any case we will use the normal distribution because we have the population standard deviation and because the sample size is "large" , i.e., above 30. The central limit theorem suggests that the sampling distribution of the sample mean approaches normal for larger samples.

b) So, now we compute a 95% $(1-\alpha)$confidence interval and see if the minimum profit, \$3,900 is inside it:

$$\overline{x} \pm z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}} \Rightarrow 4{,}110 \pm 1.96\frac{800}{\sqrt{62}} \Rightarrow (3911 \sim 4309)$$

and it turns out to be below the interval.

Here's a graph:

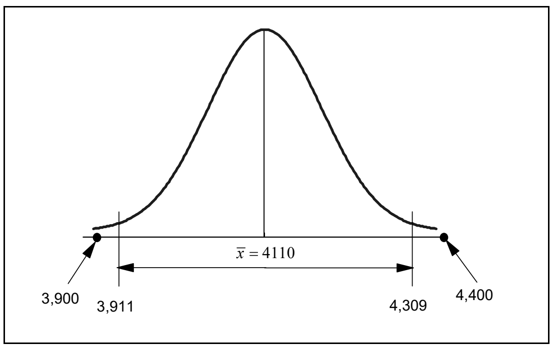


The student could also do a lower one-sided confidence interval with the smaller sample size, and that would lead to:

$$\overline{x} - z_{\alpha}\frac{\sigma}{\sqrt{n}} \Rightarrow 4{,}110 - 1.645\frac{800}{\sqrt{44}} \Rightarrow 3{,}911$$

again, not within the interval.

c) The dealer's claimed profit of $4,400 is *not* within the 95% confidence interval. The sketch looks like this:

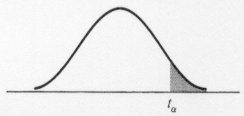


Now, note: if the student had used a small sample size of 44 and done the lower confidence interval test on the minimum profit, $3,900, he/she cannot have used the same confidence interval to test the *upper* value of $4,400, because the confidence interval is now a 2-tail interval with 5% of the area in *each* tail. That means that the confidence interval is, in fact, a 90% confidence interval, not a 95% confidence interval. So, give credit for the one-end approach in parts a) and b) but take off points in part c) if the student used the smaller sample size and the same confidence interval as in part b).

There is yet a third way the student might have proceeded; he/she might have used the smaller sample size (44) but have computed a new 95% confidence interval. That would have given:

$$\bar{x} \pm z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \Rightarrow 4{,}110 \pm 1.96 \frac{800}{\sqrt{44}} \Rightarrow (3873 \sim 4346)$$

This confidence interval would have incorrectly had the minimum profit ($3,900) in the confidence interval but the claimed profit ($4,400) would still have been outside the interval as it was supposed to be. In this case the sample size of 44 is incorrectly used with a 95% confidence interval based on that size.



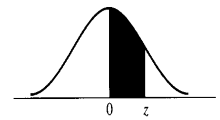
d.f.	$t_{.100}$	$t_{.050}$	$t_{.025}$	$t_{.010}$	$t_{.005}$	d.f.
1	3.078	6.314	12.706	31.821	63.657	1
2	1.886	2.920	4.303	6.965	9.925	2
3	1.638	2.353	3.182	4.541	5.841	3
4	1.533	2.132	2.776	3.747	4.604	4
5	1.476	2.015	2.571	3.365	4.032	5
6	1.440	1.943	2.447	3.143	3.707	6
7	1.415	1.895	2.365	2.998	3.499	7
8	1.397	1.860	2.306	2.896	3.355	8
9	1.383	1.833	2.262	2.821	3.250	9
10	1.372	1.812	2.228	2.764	3.169	10
11	1.363	1.796	2.201	2.718	3.106	11
12	1.356	1.782	2.179	2.681	3.055	12
13	1.350	1.771	2.160	2.650	3.012	13
14	1.345	1.761	2.145	2.624	2.977	14
15	1.341	1.753	2.131	2.602	2.947	15
16	1.337	1.746	2.120	2.583	2.921	16
17	1.333	1.740	2.110	2.567	2.898	17
18	1.330	1.734	2.101	2.552	2.878	18
19	1.328	1.729	2.093	2.539	2.861	19
20	1.325	1.725	2.086	2.528	2.845	20
21	1.323	1.721	2.080	2.518	2.831	21
22	1.321	1.717	2.074	2.508	2.819	22
23	1.319	1.714	2.069	2.500	2.807	23
24	1.318	1.711	2.064	2.492	2.797	24
25	1.316	1.708	2.060	2.485	2.787	25
26	1.315	1.706	2.056	2.479	2.779	26
27	1.314	1.703	2.052	2.473	2.771	27
28	1.313	1.701	2.048	2.467	2.763	28
29	1.311	1.699	2.045	2.462	2.756	29
inf.	1.282	1.645	1.960	2.326	2.576	inf.

Source: From "Table of Percentage Points of the *t*-Distribution." *Biometrika*, Vol. 32 (1941), p. 300. Reproduced by permission of the Biometrika Trustees.

$$\Pr\left(-t_{\alpha/2} < \frac{\bar{x}-\mu}{s/\sqrt{n}} < t_{\alpha/2}\right) = 1-\alpha$$

$$z = \frac{\hat{p}-p_0}{\sqrt{\dfrac{p_0(1-p_0)}{n}}}$$

Normal Curve Areas



z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.0000	.0040	.0080	.0120	.0160	.0199	.0239	.0279	.0319	.0359
0.1	.0398	.0438	.0478	.0517	.0557	.0596	.0636	.0675	.0714	.0753
0.2	.0793	.0832	.0871	.0910	.0948	.0987	.1026	.1064	.1103	.1141
0.3	.1179	.1217	.1255	.1293	.1331	.1368	.1406	.1443	.1480	.1517
0.4	.1554	.1591	.1628	.1664	.1700	.1736	.1772	.1808	.1844	.1879
0.5	.1915	.1950	.1985	.2019	.2054	.2088	.2123	.2157	.2190	.2224
0.6	.2257	.2291	.2324	.2357	.2389	.2422	.2454	.2486	.2517	.2549
0.7	.2580	.2611	.2642	.2673	.2704	.2734	.2764	.2794	.2823	.2852
0.8	.2881	.2910	.2939	.2967	.2995	.3023	.3051	.3078	.3106	.3133
0.9	.3159	.3186	.3212	.3238	.3264	.3289	.3315	.3340	.3365	.3389
1.0	.3413	.3438	.3461	.3485	.3508	.3531	.3554	.3577	.3599	.3621
1.1	.3643	.3665	.3686	.3708	.3729	.3749	.3770	.3790	.3810	.3830
1.2	.3849	.3869	.3888	.3907	.3925	.3944	.3962	.3980	.3997	.4015
1.3	.4032	.4049	.4066	.4082	.4099	.4115	.4131	.4147	.4162	.4177
1.4	.4192	.4207	.4222	.4236	.4251	.4265	.4279	.4292	.4306	.4319
1.5	.4332	.4345	.4357	.4370	.4382	.4394	.4406	.4418	.4429	.4441
1.6	.4452	.4463	.4474	.4484	.4495	.4505	.4515	.4525	.4535	.4545
1.7	.4554	.4564	.4573	.4582	.4591	.4599	.4608	.4616	.4625	.4633
1.8	.4641	.4649	.4656	.4664	.4671	.4678	.4686	.4693	.4699	.4706
1.9	.4713	.4719	.4726	.4732	.4738	.4744	.4750	.4756	.4761	.4767
2.0	.4772	.4778	.4783	.4788	.4793	.4798	.4803	.4808	.4812	.4817
2.1	.4821	.4826	.4830	.4834	.4838	.4842	.4846	.4850	.4854	.4857
2.2	.4861	.4864	.4868	.4871	.4875	.4878	.4881	.4884	.4887	.4890
2.3	.4893	.4896	.4898	.4901	.4904	.4906	.4909	.4911	.4913	.4916
2.4	.4918	.4920	.4922	.4925	.4927	.4929	.4931	.4932	.4934	.4936
2.5	.4938	.4940	.4941	.4943	.4945	.4946	.4948	.4949	.4951	.4952
2.6	.4953	.4955	.4956	.4957	.4959	.4960	.4961	.4962	.4963	.4964
2.7	.4965	.4966	.4967	.4968	.4969	.4970	.4971	.4972	.4973	.4974
2.8	.4974	.4975	.4976	.4977	.4977	.4978	.4979	.4979	.4980	.4981
2.9	.4981	.4982	.4982	.4983	.4984	.4984	.4985	.4985	.4986	.4986
3.0	.4987	.4987	.4987	.4988	.4988	.4989	.4989	.4989	.4990	.4990

Source: This table is abridged from Table 1 of *Statistical Tables and Formulas*, by A. Hald (New York: John Wiley & Sons, Inc., 1952). Reproduced by permission of A. Hald and the publishers, John Wiley & Sons, Inc.

$$n = \frac{z_{\alpha/2}^2\sigma^2}{D^2}. \qquad n = \frac{z_{\alpha/2}^2}{4D^2}$$

Binomial Coefficients

n	$\binom{n}{0}$	$\binom{n}{1}$	$\binom{n}{2}$	$\binom{n}{3}$	$\binom{n}{4}$	$\binom{n}{5}$	$\binom{n}{6}$	$\binom{n}{7}$	$\binom{n}{8}$	$\binom{n}{9}$	$\binom{n}{10}$
0	1										
1	1	1									
2	1	2	1								
3	1	3	3	1							
4	1	4	6	4	1						
5	1	5	10	10	5	1					
0	1	6	15	20	15	6	1				
7	1	7	21	35	35	21	7	1			
8	1	8	28	56	70	56	28	8	1		
9	1	9	36	84	126	126	84	36	9	1	
10	1	10	45	120	210	252	210	120	45	10	1
11	1	11	55	165	330	462	462	330	165	55	11
12	1	12	66	220	405	792	924	792	495	220	66
13	1	13	78	286	715	1287	1716	1716	1287	715	286
14	1	14	91	364	1001	2002	3003	3432	3003	2002	1001
15	1	15	105	455	1365	3003	5005	6435	6435	5005	3003
16	1	16	120	560	1820	4368	8008	11440	12870	11440	8008
17	1	17	136	680	2380	6188	12376	19448	24310	24310	19448
18	1	18	153	816	3060	8568	18564	31824	43758	48620	43758
19	1	19	171	969	3876	11628	27132	50388	75582	92378	92378
20	1	20	190	1140	4845	15304	38760	77520	125970	167960	184756

If necessary, use the identity $\binom{n}{k}=\binom{n}{n-k}$.

$$f(\tilde{x})=\frac{1}{\sigma\sqrt{2\pi}}e^{-(x-\mu)^2/2\sigma^2}\quad -\infty<x<\infty$$

$$\sigma^2=\frac{\sum_{i=1}^{N}(x_i-\mu)^2}{N}$$

$$p(x)=\frac{C_x^r C_{n-x}^{N-r}}{C_n^N}=\frac{\binom{r}{x}\binom{N-r}{n-x}}{\binom{N}{n}}$$

Mean: $\mu = n\left(\frac{r}{N}\right)$

Variance: $\sigma^2 = n\left(\frac{r}{N}\right)\left(\frac{N-r}{N}\right)\left(\frac{N-n}{N-1}\right)$

Standard deviation: $\sigma = \sqrt{\sigma^2}$

$$f(x)=\begin{cases}\dfrac{1}{(b-a)}, & a\le x\le b\\ 0, & \text{otherwise}\end{cases}$$

$$\mu=\frac{1}{2}(b+a)\text{ and }\sigma=\frac{(b-a)}{\sqrt{12}}$$

$$f(x)=\begin{cases}\lambda e^{-\lambda x}, & \lambda>0, x\ge 0\\ 0, & \text{otherwise}\end{cases}$$

$$\mu=\frac{1}{\lambda}\text{ and }\sigma=\frac{1}{\lambda}$$

$$P(x\ge a)=e^{-\lambda a},\ a\ge 0\ and\ \lambda>0$$

$$P(A_i|B)=\frac{P(B|A_i)P(A_i)}{\sum_{\text{all }k}P(B|A_k)P(A_k)}$$

$$P(x)=\begin{cases}\dfrac{e^{-\lambda t}(\lambda t)^x}{x!}, & \text{for } x=0,1,2,,,\infty,\quad \lambda>0,\\ 0, & \text{otherwise.}\end{cases}$$

λ = the mean number of events in a given segment of time ($t=1$)

t = the length of a particular subsegment ($t\le 1$)

$E[x]=\mu_x=\lambda t$ = the expected number of events in one subsegment length t