Statistics Test 7 Review

Chapters 11, 12, 14 and 15

What you should know (use your textbook to look these up):

- What a simulation is
- How to generate (or simulate) random numbers
- How to use a random number table
- Difference among a population, a sample frame and a sample.
- Difference between parameter and statistic
- What bias is, and types of bias: nonresponse bias, voluntary response bias, response bias
- Types of samples: SRS, stratified sample, cluster sample, systematic sample, convenience sample, multistage sample
- What a census is
- The Law of Large Numbers
- Sample space
- The sum of probabilities of elementary outcomes is 1
- The complement rule
- Disjoint (or mutually exclusive) events
- Addition rule
- Independence of events
- Multiplication rule
- Conditional probability

Practice Problems:

1. There are eight different car keys in a drawer, including yours. Suppose you grab one key at a time until you get your car key. Estimate the probability that you get your car key on the second try. Describe the method that you will use to simulate this exercise. Use the attached random number table to do your estimation.

   This is an ***example*** *of an answer. It is not the only possible answer out there.*
   I'm going to assign the following single digits as follows:
   1: my key
   2 – 8: other keys
   9, 0: ignore

   I'm then going to go through the table, selecting pairs of allowable (not 9 or 0) digits. Those that are doubles (like 11, 22, etc) I will also ignore, since I wouldn't select the same key twice. I will count the ones with a 1 in the second digit as successes, and then I'll estimate the probability by dividing my successes by the number of trials.

   I did this using the random number table (see last page), and I got a total of 7 out of 59, or about 12%.

2. A statistics teacher wants to know how her students feel about an introductory course. She decides to administer a survey to a random sample of students taking the course. She has several sampling plans to choose from. Name the sampling strategy in each, and comment on the strategy (whether it's a good strategy or not, and why)

   a) There are four classes of students taking the course: freshmen, sophomores, juniors and seniors. Randomly select 15 students from each class rank.
   Stratified sample– it is a generally accepted method, and as long as there are about the same number of each class (freshman, etc.) in the course, then it ensures that no particular class is underrepresented. If there are far more freshmen than seniors, then she should have chosen more freshmen than seniors for her sample.

   b) Each student has a 9-digit student number. Put the numbers in order, and with a random number generator, randomly select 60 numbers.
   SRS. The best method, and the one that our statistical methods assume that is being used. Every combination of students has the same chance of being chosen as any other.

   c) Randomly select a class (freshmen, sophomores, juniors and seniors) and survey every student in that class.
   Cluster sample. In this particular case this seems like a not very smart choice, since it ensures that only one class will be represented. If there are differences in opinion among classes, this could give you a very non-representative sample.

   d) Using the class roster, select every fifth student from the list, starting from a randomly chosen person.
   Systematic sample. It's generally considered a good sampling method.

   e) Offer extra credit to students to volunteer for the survey.
   This is just a terrible idea. Only students who are in need of extra credit will volunteer. Voluntary response bias and non-response bias galore.

   f) At the beginning of class one day, survey everyone present.
   Also a bad idea, because you're only getting those who make it to class on time. There's a good chance that those who are unhappy with the class won't make it to class on time, which means that your sample will be biased toward a positive response. If you're looking for a positive response rather than a sample that truly reflects the whole class then maybe this is a good method, but if you're a self respecting statistician you'll choose something else.

3. Suppose a state has 10 universities, 25 four-year colleges, and 50 community colleges, each of which offers multiple sections of an introductory statistics class each year. Researchers want to conduct a survey of 100 students taking introductory statistics in the state. Explain a method for collecting each of the following types of samples:
   a) A **stratified sample**. b) A **cluster sample**. c) A **simple random sample**.

   Again, this is an ***example*** *of an answer. It is by no means the only answer:*

   a) Stratified sampling could be designed as follows: Get a current list of all students enrolled in statistics at the 85 institutions, as well as the type of the institution in which they are enrolled. Calculate the proportions of students in the universities, the 4-year colleges and the community colleges. Using the same proportions, randomly select the appropriate number of students from each *type of school. For example, if 15% of the students are in universities, 40% from 4-year colleges and 45% from community colleges, then take a simple random sample of 15 students from the list of university students, 40 from the list of 4-yr college students and 45 from the list of community college students.*

b) Cluster sampling could be designed as follows: Get a current list of all intro stats courses, and randomly select 5 of those classes. Survey all students in each class.

c) SRS could be designed as follows. Get a current list of all students enrolled in statistics at the 85 institutions. Number all students and randomly (using a random number generator, for example) select 100 students from the entire population.

4. What are some advantages/disadvantages of each of the prior question's data collection methods?
a) Stratified—Ensures that each population (university, 4-yr college and community college) is proportionally represented, reducing bias that might occur from overrepresentation of one of the types of schools. It is not as statistically sound an approach as the SRS, but it's good.
b) Clustering—is more feasible and may give a larger sample of data to work from. The downside is that it might also overrepresent some parts of the population and underrepresent others.
c) SRS—is the best approach but not particularly feasible or practical. This is the approach that our inferential statistics methods are based on, and is unfortunately the hardest one to actually pull off.

5. Which type of bias (if any) do you think would be introduced in each of the following situations?

a) A list of registered automobile owners is used to select a random sample for a survey about whether people think homeowners should pay a surtax to support public parks.
Non-response bias, unless they can get a lot of people to respond to the survey. If a small number respond, then it will probably be biased toward those who are angry by the idea of having their taxes pay for something they might not use. And then the sampling frame is poorly chosen: the sampling frame is all registered automobile owners. This list is going to include a greater percentage of homeowners than there is in the general population, since people who don't own a car frequently don't own their home either. This means that homeowners will likely be overrepresented, and those that do not own their home will be underrepresented. This is similar to the Literary Digest scenario that predicted Alf Landon would defeat Roosevelt in the '36 presidential election.

b) A survey is mailed to a random sample of residents in a city asking if they think the current mayor is doing an acceptable job. *Non-response bias—mailed surveys get a pretty poor response rate. Voluntary reponse bias—only those who choose to respond will, and those will be the ones with strong feelings. Possible response bias—asking if the mayor is doing an "acceptable" job is somewhat negatively worded—after all, the choices are "acceptable" (sounds mediocre) versus "not acceptable" (sounds really bad). An open ended question would be less leading.*

c) In a college town, college students are hired to conduct door-to-door interviews, based on a multistage cluster sample, to determine if city residents think there should be a law forbidding loud music at parties. *Definite response bias. The college kids are conducting the interviews, and they're also the ones throwing the parties with loud music. The respondents will be less likely to voice support for a law banning loud music if they're talking to the students than if they're talking to a neutral interviewer.*

d) A magazine sends a survey to a random sample of its subscribers asking them if they would like the frequency of publication reduced from biweekly to monthly, or would prefer that it remain the same.

Non-response bias, since they are not likely to get a very good response here. But I think the real problem here is the sampling frame. What is the population in question: is it all subscribers to the magazine, or is it all potential subscribers to the magazine? If they think that they might lure more subscribers by changing the frequency of publication, then their sampling frame is poorly chosen, because they are only asking people that already subscribed with the current frequency of publication.

6. Gastwirth (1988, p. 507) describes a court case in which Bristol Myers was ordered by the Federal Trade Commission to stop advertising that "twice as many dentists use Ipana as any other dentifrice" and that more dentists recommended it than any other dentifrice. Bristol Myers based its claim on a survey of 10,000 randomly selected dentists from a list of 66,000 subscribers to two dental magazines. They received 1983 responses, with 621 saying they used Ipana and only 258 reporting that they used the second most popular brand. As for the recommendations, 461 respondents recommended Ipana, compared with 195 for the second most popular choice.

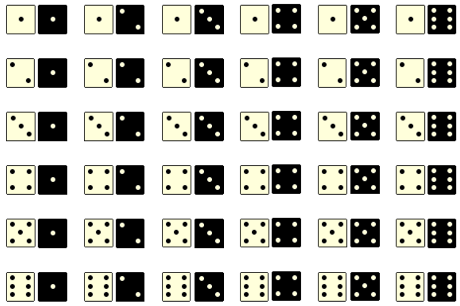
   a) Specify the **sampling frame** for this survey, and explain whether or not you think "using the wrong sampling frame" was a problem here, based on what Bristol Myers was trying to conclude.
   A sampling frame is the list of units from which the sample is selected. In this case it is the subscribers from the two magazines. The population consists of all dentists. There was probably a large sector of dentists not surveyed, and if that is the case, then yes, the sample frame was a problem here.
   b) What could Bristol Myers have done to improve the validity of the results after it had mailed the 10,000 surveys and received only 1983 of them back? (Assume it kept track of who had responded and who had not.) *They could have followed up the non-responses with phone calls.*

   By the way, this court case occurred in 1950, and it referred to a survey done in 1940, which is why you've probably never heard of that particular brand of toothpaste.

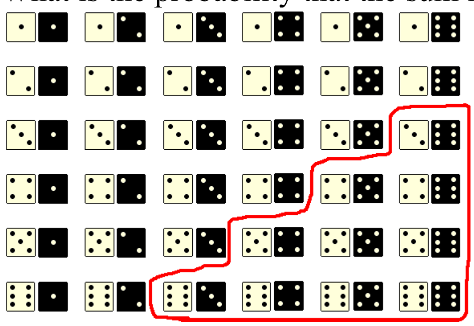
7. An Internet poll sponsored by a site called About.com asked Internet users to pick one of two choices in response to the question, "Should jurors opposed to gun control laws refuse to convict defendants even if they have clearly broken the laws?" The two choices and the number and percent choosing them were
   • Yes, that's an effective way to defeat unjust laws (16,864, 23%)
   • No, that undermines the legal system (55,519, 77%)

   Discuss this poll, including whether or not you think the results are **representative** of all adults and whether you think the wording is appropriate.
   The sampling frame is poor: the population is "all adults" and the sampling frame is those people (adults and youth alike) that have internet access and have come across this poll. And then the survey suffers from voluntary response bias as well: only those who choose to (because they have strong feelings about the subject) will respond at all.
   One might also have a problem with the wording. The statement that it is an "effective way to defeat unjust laws" implies that gun control laws are unjust. It could be worded more neutrally than it is.

8. You have rolled two fair six-sided dice.
   a) List the sample space for the **sum** of two rolled fair dice.
      {2, 3, 4, 5, 6, 7, 8, 9,10, 11, 12}
      Note: this is the sample space for the sums. It is not the sample space for the ROLLS, which would be the 36 pairs:



b) What is the probability that the sum is greater than 8?



$$P(\text{sum} > 8) = \frac{\text{number of possible rolls where sum} > 8}{\text{number of all possible rolls}} = \frac{10}{36} \approx .278$$

c) What is the probability that the sum is greater than or equal to 8?

$$P(\text{sum} \geq 8) = P(\text{sum} > 8) + P(\text{sum} = 8) = \frac{10}{36} + \frac{5}{36} = \frac{15}{36} = \frac{5}{12} \approx .417$$

d) What is the probability that the sum is odd?

$$P(\text{sum is odd}) = \frac{18}{36} = \frac{1}{2} = .5$$

e) What is the probability that the sum is odd and even?
It can't be both. P(both odd and even) = 0.
Odd and even are called mutually exclusive, or disjoint events.

f) What is the probability that the sum is greater than 12?
All possible sums are 12 or less, so P(sums > 12) = 0

g) What is the probability that the sum is greater than 9 GIVEN that the sum is even?

$$P(\text{sum} > 9 \,|\, \text{sum is even}) = \frac{P(\text{sum} > 9 \text{ and is even})}{P(\text{sum is even})}$$

$$= \frac{P(\text{sum} = 10 \text{ or sum} = 12)}{P(\text{sum is even})} = \frac{\frac{3}{36} + \frac{1}{36}}{\frac{1}{2}} = \frac{4}{18} = \frac{2}{9} \approx .222$$

h) What is the probability that the sum is odd GIVEN that the sum is greater than 6?

$$P(\text{sum is odd} \mid \text{sum} > 6) = \frac{P(\text{sum is odd and} > 6)}{P(\text{sum} > 6)} = \frac{\frac{6}{36}}{\frac{21}{36}} = \frac{6}{21} = \frac{2}{7} \approx .286$$

i) If A is an even roll and B is a roll greater than 9, are events A and B dependent or independent? Why? Are they disjoint or not? Why?

$P(\text{sum} > 9 \mid \text{sum is even}) \approx .222$ *(from part g above).*

$P(\text{sum} > 9) = \frac{6}{36} = \frac{1}{6} \approx .167$. *Since the conditional probability is not equal to the marginal probability, the events are not independent. In other words, when we learn that the sum is even, the probability that it is >9 goes down from 22.2% to 16.7%. If they were independent, the probability would be unchanged.*

The events are not disjoint. If they were, then $P(\text{sum} > 9 \mid \text{sum is even})$ *would be zero.*

9. Home security experts estimate that an untrained housedog has a 70% chance of detecting an intruder. They also believe that the dog has a 50% of scaring the intruder away IF he/she has been detected (undetected intruders don't get scared away). Assume "Fido" is an untrained housedog; what is the probability that Fido will successfully scare away an intruder?

$$P(\text{detects and scares intruder}) = P(\text{scares intruder} \mid \text{detects intruder}) \cdot P(\text{detects intruder})$$
$$= (.5)(.7) = 35\%$$

10. Two questions were asked of 50 randomly selected football fans at the UT-A&M game. (1) What team were they cheering for? And (2) were they wearing shoes? Let A be the event of being an A&M fan. Let S be the event of wearing shoes.

	UT fans	A&M fans
Wearing shoes	4	30
Not wearing shoes	16	0

a) Find $P(A)$.

$P(A) = \frac{30}{50} = .6$

b) Find $P(S)$.

$P(S) = \frac{34}{50} = .68$

c) Find $P(A$ and $S)$.

$P(A \cap S) = \frac{30}{50} = .6$

d) Find $P(A$ or $S)$.

$P(A \cup S) = \frac{34}{50} = .68$

e) Find $P(A \mid S)$.

$P(A \mid S) = \frac{30}{34} \approx .882$

f) Find $P(S \mid A)$.

$P(S \mid A) = \frac{30}{30} = 1$

11. A card is selected from a well-shuffled deck of 52 cards. Let H be the event that the card is a heart, and let F be the event that the card is a face card (King, Queen, Jack).

a) Determine $P(H)$.

There are 13 hearts in a deck of cards, so $P(H) = \frac{13}{52} = .25$

b) Determine $P(F)$.

There are 12 face cards (3 in each suit) in a deck of cards, so $P(H) = \frac{12}{52} \approx .231$

c) Determine $P(H \mid F)$.

You can look at this two ways:

How many face cards are hearts? Answer: 3 out of 12, so P(H | F) = .25

Or, you can use the formula: $P(H \mid F) = \frac{P(H \text{ and } F)}{P(F)} = \frac{3/52}{12/52} = \frac{3}{12} = .25$

d) Determine $P(F \mid H)$.

Again, you can look at this two ways:

How many hearts are face cards? Answer: 3 out of 13, so P(H | F) ≈ .231

Or, you can use the formula: $P(F \mid H) = \frac{P(H \text{ and } F)}{P(H)} = \frac{3/52}{13/52} = \frac{3}{13} \approx .231$

e) Are H and F independent events? Explain intuitively (or using parts (b) and (d)) your answer regarding independence.

The intuitive answer is that the suit of the card does not affect whether it's a face card or not, so the two events are independent. The mathematical answer is that since P(H | F) = P(H) and P(F | H) = P(F), the events are independent.

12. At a fictional high school Statistics High, after-school activities can be classified into three types: athletes, fine arts, and other. The table gives the number of students participating in each of these types of activities by grade. For the purpose of this example, we will assume that any given student is in exactly one of these after-school activities.

	9th	**10th**	**11th**	**12th**	**Total**
Athletes	150	160	140	150	600
Fine Arts	100	90	120	125	435
Other	125	140	150	150	565
Total	375	390	410	425	1600

a) What is the probability that a student selected is a 9th grader? $P(\text{9th grader}) = \frac{375}{1600} \approx 23.4\%$

b) What is the probability that a student selected is an athlete? $P(\text{athlete}) = \frac{600}{1600} \approx 37.5\%$

c) What is the probability that a student selected is a junior or senior?

$P(\text{11th or 12th}) = \frac{410 + 425}{1600} \approx 52.2\%$

d) Are being an athlete and being in the 11^{th} grade dependent or independent events?

$P(\text{athlete} \mid \text{11th grade}) = \frac{140}{410} \approx 34.1\%$. *Since P(athlete) ≠ P(athlete | 11th grade), the events are not independent. (Close, but not quite)*

e) What is the probability that a student selected is a 9th grader, given the student selected is an athlete? $P(\text{9th grade} \mid \text{athlete}) = \frac{150}{600} = 25\%$.

f) What is the probability that a student selected is an athlete, given the student selected is a 9th grader? $P(\text{athlete} \mid \text{9th grade}) = \frac{150}{375} = 40\%$.

g) What is the probability that a student selected is a junior or senior, given the student is in fine arts? $P(\text{junior or senior} \mid \text{fine arts}) = \frac{245}{435} = 56.3\%$.

h) What is the probability that a student selected is in fine arts, given the student is a junior or senior? $P(\text{fine arts} \mid \text{junior or senior}) = \frac{245}{835} = 29.3\%$.

13. An insurance company has the following information about drivers aged 16 to 18 years: 20% are involved in accidents each year; 10% in this age group are A students; among those involved in accidents, 5% are A students.
I'm making a table to help myself here:
10% of all 16 to 18 yr olds are A students. 20% of all 16 to 18yr olds are in accidents. Those are both marginal probabilities.
Since 5% of the kids in accidents are A students, that means that they are 5% of 20% of the total, meaning 1% of the total age group are both A students and in accidents. Now I can fill in the rest of the table by making the rows and columns add up appropriately:

	A student	Not A student	Total
In an accident	1%	19%	20%
Not in an accident	9%	71%	80%
Total	10%	90%	100%

a) What is the probability that a driver from this age group is involved in an accident?
$P(\text{in an accident}) = 20\%$

b) What is the probability that a driver that is not involved in an accident is not an A student?

$$P(\text{not an A student} \mid \text{not in an accident}) = \frac{P(\text{not an A student and not in an accident})}{P(\text{not in an accident})}$$

$$= \frac{.71}{.8} = 88.75\%$$

c) What is the probability that an A student driver is involved in an accident?

$$P(\text{in an accident} \mid \text{A student}) = \frac{P(\text{an A student and in an accident})}{P(\text{an A student})}$$

$$= \frac{.01}{.1} = 10\%$$

14. I've got a little leather pouch with some coins in it. I've got 3 quarters, 8 nickels, 4 dimes and a penny. I start drawing coins randomly from the pouch.

a) What is the probability that the first two coins are quarters?

The probability that the first coin is a quarter is $\frac{3}{16}$*, since we have 3 quarters out of 16 total coins. Next we need to look at the* ***conditional probability****:*

P(second coin is a quarter | first coin is a quarter) = $\frac{2}{15}$*, since if we have already picked a quarter out of the pouch, we have one less quarter and one less coin than we had before. So the probability of both getting a quarter on the first pick and a quarter on the second pick is:*

P(first is a quarter) × P(second is a quarter | first is a quarter) = $\frac{3}{16} \cdot \frac{2}{15} = \frac{1}{40} =$ 2.5%.

b) What is the probability that the first four coins are quarters?

Since we only have 3 quarters, this cannot happen. P(four quarters) = 0.

c) What is the probability that the value of the first two coins picked is more than 30 cents?

OK, the only ways we can get more than 30 cents is by choosing two quarters or a quarter and a dime. So the possible picks would be: QQ, QD, or DQ. So the probability that we're going to get at least 30 cents is:

P(first is quarter) × P(second is either dime or quarter | first is quarter)

+ P(first is dime) × P(second is quarter | first is dime)

$$= \frac{3}{16} \cdot \frac{6}{15} + \frac{4}{16} \cdot \frac{3}{15} = \frac{18 + 12}{16 \cdot 15} = \frac{30}{16 \cdot 15} = \frac{1}{8}$$

d) What is the probability that in the first 5 coins picked, one of them is a penny?

This is one of those where it's easier to use the complement rule: let's calculate the probability that ***none*** *of the first 5 coins picked is a penny. The probability that the first one isn't a penny is* $\frac{15}{16}$*, and the probability that the second one isn't a penny (given that the first one isn't a penny) is* $\frac{14}{15}$*, and so on for five picks. So we get that the probability of getting no pennies in the first 5 picks is:* $\frac{15}{16} \cdot \frac{14}{15} \cdot \frac{13}{14} \cdot \frac{12}{13} \cdot \frac{11}{12} = \frac{11}{16}$*. This tells us that the probability of getting a penny in the first 5 picks is* $1 - \frac{11}{16} = \frac{5}{16}$.

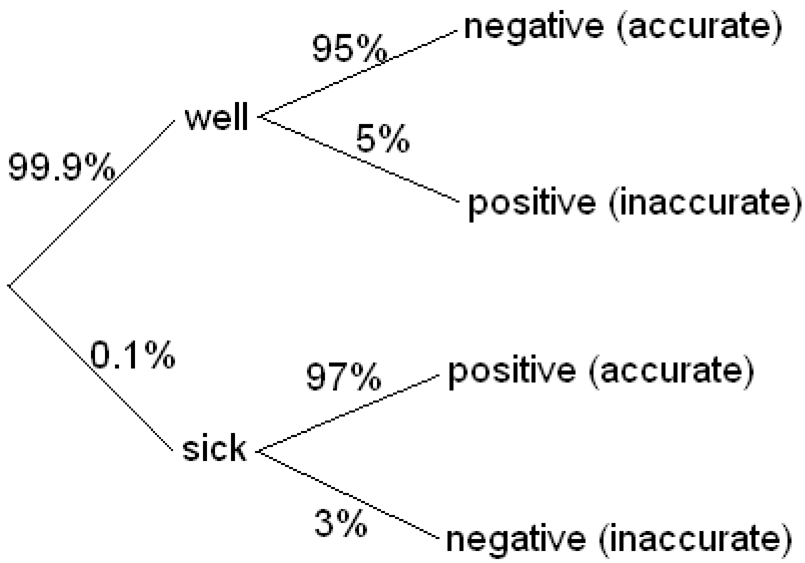
e) What is the probability that in the first 5 coins picked, at least one of them is a nickel?

Use the same strategy as the prior problem. P(no nickels in first 5 coins picked) = $\frac{8}{16} \cdot \frac{7}{15} \cdot \frac{6}{14} \cdot \frac{5}{13} \cdot \frac{4}{12} = \frac{1}{78}$*. So the probability of getting at least one nickel is* $\frac{77}{78}$.

Bonus question –the one that I started to go over in class, but we ran out of time:

Suppose there is an outbreak of a rare disease in the United States. It is estimated that 1 in every 1,000 people will contract the disease. A diagnostic test is available for this disease that is 97% accurate for people who have the disease and 95% accurate for people who do not have the disease. A randomly selected group of people is given the test.
a. What is the probability that someone who tested positive actually has the disease?

OK, first define the events:
The first event is sick versus well, and the second event is positive versus negative result. You could also define the second event to be accurate or inaccurate result. Let's draw a probability tree:



So if we take these results, multiply the probabilities and put them into a table, we get the following:

	Well	Sick	Total
Positive	0.04995	0.00097	0.05092
Negative	0.94905	0.00003	0.94908
Total	0.99900	0.00100	1.00000

We can now see that $P(\text{sick} \mid \text{tested positive}) = \dfrac{P(\text{sick and positive})}{P(\text{tested positive})} = \dfrac{.00097}{.05092} \approx 2\%$.

This gives us the bizarre (but correct) answer that only 2% of those testing positively actually have the disease. The reason behind this bizarre result is that the disease is so rare. It is much rarer to have the disease than it is to have a false positive test. So if you have a positive test, you should not necessarily freak out. There's still a 98% probability that you're fine.

b. What is the probability that someone who tested negative actually has the disease?

$$P(\text{sick} \mid \text{tested negative}) = \frac{P(\text{sick and negative})}{P(\text{tested negative})} = \frac{.00003}{.94908} \approx .003\%.$$

This gives us the much more expected answer that if you get a negative result you should feel pretty good about your chances. More than pretty good, actually. Really confident.

69571	44927	06220	71020	78336	38877	00512	99602	0 out of 14
44617	41152	51398	10533	86731	97824	37955	48185	3 out of 16
99999	88276	18346	58650	56476	86504	42111	33055	1 out of 12
57553	52822	73125	08463	85731	71003	38491	83871	3 out of 17
23261	56141	19828	85055	75202	35032	63432	23671	
18235	61204	20160	14732	53790	71839	46288	65321	
44730	36327	36584	87768	81319	82043	63202	87389	
84000	45607	41131	53658	34674	04314	75604	62243	
67164	20190	04210	31468	39707	20324	10213	68265	
43346	54771	61884	21828	73997	00342	87399	64070	
32085	41043	67095	00345	90652	58516	90238	70236	
86442	28429	97092	09278	64899	28093	01044	88919	
92141	68958	35269	45256	27134	85131	69970	60972	
59575	64434	17530	20336	33251	80109	70113	72821	
98612	86304	84321	41296	62043	36671	52437	58522	
51943	81752	07875	79525	81825	88449	79106	41168	
69403	66930	90604	41676	70485	94858	75185	83296	
43815	41340	35931	77408	44745	74287	95330	20441	
08964	05563	58535	26733	12657	39700	25477	51730	
80507	94365	85305	71200	01490	73697	84489	41164	
14876	82436	46119	46691	95955	59889	41813	12080	
30921	16762	12114	19812	42948	44588	71891	64129	
55284	72329	43629	20608	29073	28534	70755	74158	
16426	19098	43442	62310	27640	30192	97182	76891	
64893	94094	46415	18223	07581	11360	54634	15035	
03886	06189	06314	10891	01285	50318	05836	97248	
80328	10512	11746	00333	64083	99318	95838	52121	
00937	67446	38101	96227	59182	21174	47943	13984	
20033	63132	33922	86796	47463	30397	84300	21911	
31802	15446	09609	58808	18845	87998	34674	88711	
10238	53709	83741	70979	20691	26281	27770	29032	
95268	31476	62524	67929	00626	43327	92862	72938	
17042	56359	01132	87950	94396	74996	77917	96424	
49408	46927	52476	99544	54157	04786	84517	65870	
12024	87918	40636	07284	49953	87265	99804	13822	
46081	98670	00166	85395	97728	27594	74548	52102	