STT 315
Preparation for the FINAL EXAM (longer than any actual exam).

Note: According to the syllabus your course grade is based upon total points. Each graded activity such as exams, homework, bonus, is converted into such points by means of a curve. The relationship of total course points to course grade is given in the syllabus.

New: Your total course points will be **the larger of**

a. Total points figured by the method outlined in the syllabus (usual method).
b. Total points figured as in (a), except that the points you earn for your lowest exam will be replaced by **(23/30) times (points you earn for the final exam)**.

1. $P(A) = 0.7$, $P(B) = 0.4$.

a. What is the very least P(AB) could be?
ans. a. P(A union B) is the same as P(A or B) and cannot exceed one. But

b. $P(A \text{ or } B) = P(A) + P(B) - P(A\ B) = 0.7 + 0.4 - P(A\ B) = 1.1 - P(AB)$

so P(AB) must be at least 0.1. If $P(AB) = 0.1$ then $P(A \text{ or } B) = 1.1 - 0.1 = 1$. For this case

c. $P(B \mid A) = P(AB) / P(A) = 0.1 / 0.7$
d. $P(AB) = 0.1$ as shown above. AB is another name for 'A and B."
e. $P(A^C) = 1 - P(A) = 1 - 0.7 = 0.3$

b. If A, B are independent events what are
a. $P(AB) = P(A)\ P(B) = 0.7\ \ 0.4 = 0.28$
b. $P(A \text{ or } B) = P(A) + P(B) - P(AB) = 0.7 + \ 0.4 - 0.28$
c. $P(B \mid A) = P(B) = 0.4$ since, for independent events, knowing A has occurred does not change the probability for B. Formally, $P(AB) / P(A) =$ (for independent events) $P(A)\ P(B) / P(A) = P(B) = 0.4$.
d. $P(B \mid A^C) = P(B) = 0.4$.

2. Let events A, B be
A = customer will buy the product
B = customer leaves a cash deposit
Suppose that $P(A) = 0.3$, $P(B \mid A) = 0.8$, $P(B \mid A^C) = 0.4$.
a. Make a complete tree diagram for this information.

$P(B|A) = 0.8$ $\quad P(AB) = P(A)\ P(B|A) = 0.3\ \ 0.8 = 0.24$
$P(A) = 0.3$
$P(B^C|A) = 0.2$ $\quad P(AB^C) = P(A)\ P(B^C|A) = 0.3\ \ 0.2 = 0.06$

$P(B|A^C) = 0.4$ $\quad P(A^CB) = P(A^C)\ P(B|A^C) = 0.7\ \ 0.4 = 0.28$
$P(A^C) = 0.7$
$P(B^C|A^C) = 0.6$ $\quad P(A^CB^C) = P(A^C)\ P(B^C|A^C) = 0.7\ \ 0.6 = 0.42$

b. Make a complete Venn diagram for this information.

[ ] <---- circle these two for B

correction----> 0.06 0.24 0.28 0.42

circle these two for A----> [ ] [ ] this is $A^C B^C$

c. Determine P(B) and also $P(B^C)$.

$P(B) = P(AB) + P(A^C B) = 0.24 + 0.28 = 0.52$ (corrected)

$P(B^C) = 1 - P(B) = 1 - 0.52 = 0.48$ (corrected)

d. Determine P(A | B) and also $P(A \mid B^C)$.

$P(A|B) = P(AB) / P(B) = 0.24 / 0.52$ (corrected)

3. $P(OIL) = 0.2$, $P(+ \mid OIL) = 0.8$, $P(+ \mid OIL^C) = 0.3$, cost to test is 40, cost to drill is 300, return from oil is 1200.

a. Make a complete tree diagram including all endpoint probabilities and the consequent net returns x from the policy "test, but drill only if the test is positive."

net

$P(+ \mid OIL) = 0.8$ $P(OIL+) = P(OIL)\,P(+ \mid OIL) = 0.2 \cdot 0.8 = 0.16$ 860

$P(OIL) = 0.2$

$P(- \mid OIL) = 0.2$ $P(OIL-) = P(OIL)\,P(- \mid OIL) = 0.2 \cdot 0.2 = 0.04$ -40

$P(+ \mid OIL^C) = 0.3$ $P(OIL^C +) = P(OIL^C)\,P(+ \mid OIL^C) = 0.8 \cdot 0.3 = 0.24$ -340

$P(OIL^C) = 0.8$

$P(- \mid OIL^C) = 0.7$ $P(OIL^C -) = P(OIL^C)\,P(- \mid OIL^C) = 0.8 \cdot 0.7 = 0.56$ -40

b. Find E X.

The net returns x, and their probabilities, for each of the above for tree endpoints are

	x	p(x)	x p(x)
– 40 – 300 + 1200 =	860	0.16	860 0.16
– 40 – 00 + 00 =	– 40	0.04	–40 0.04
– 40 – 300 + 00 =	– 340	0.24	–340 0.24
– 40 – 00 + 00 =	– 40	0.56	–40 0.56

E X = sum of x p(x) = 32

c. Find E Y where y is the net return from the policy "just drill, do not test."

E Y = 0.2 (1200 – 300) + 0.8 (–300) = –60.

4. Drawing balls with equal probability but without replacement from {R R G G G Y}.

a. P(Y2) (guess it from a principle that you properly name and confirm your guess using the rules of total probability and multiplication). Identify your use of the rules.

Order of the deal does not matter so P(Y2) = P(Y1) = 1/6. Using total probability we have $P(Y2) = P(Y1\ Y2) + P(Y1^C\ Y2) = 0 + P(Y1^C\ Y2) = P(Y1^C)\,P(Y2 \mid Y1^C)$

$= 5/6 \cdot 1/5 = 1/6$

b. P(G1 | Y2). Prove your answer using the rules.
P(G1 | Y2) is the same as P(G2 | Y1) = 3 / 5 since order of the deal does not matter. To prove it using the rules use P(G1 | Y2) = P(G1 Y2) / P(Y2) where

P(G1 Y2) = P(G1) P(Y2 | G1) = 3/6 1/5

P(Y2) = P(Y1) = 1/6

So P(G1 | Y2) = (3/6 1/5) / (1/6) = 3/5, confirming the fact that order of the deal does not matter.

5. P(rain today) = 0.6, P(rain tomorrow) = 0.5, P(rain tomorrow | rain today) = 0.8.
a. P(rain today and rain tomorrow).
P(tod tom) = P(tod) P(tom | tod) = 0.6 0.8 = 0.48

b. Give the complete Venn diagram.

tod tom 0.6 0.8 = 0.48
tod tom^C 0.6 (1 – 0.8) = 0.12
using the decomposition $tod^C tom$ = tom – tod tom (draw a picture of this!)
$tod^C tom$ P(tod) – P(tod tom) = 0.5 – 0.6 0.8 = 0.02
$tod^C tom^C$ 1 – sum of above = 1 – (0.48 + 0.12 + 0.02) = 0.38

c. Give the complete tree diagram.

P(tom | tod) = 0.8 P(tod tom) = 0.6 0.8 = 0.48
P(tod) = 0.6
P(tom^C | tod) = 0.2 P(tod tom^C) = 0.6 0.2 = 0.12

P(tom | tod^C) = 0.02 / 0.4 P($tod^C tom$) = 0.5 – 0.6 0.8 = 0.02
P(tod^C) = 0.4
P(tom^C | tod^C) = 0.38 / 0.4 P($tod^C tom^C$) = 1 – 0.62 = 0.38

d. Give P(rain today | rain tomorrow).
P(tod | tom) = P(tod tom) / P(tom) = 0.48 / 0.5

6.

	x	p(x)	x p(x)	x^2 p(x)
	0	0.2	0 0.2 = 0.0	0 0.2 = 0.0
	1	0.8	1 0.8 = 0.8	1 0.8 = 0.8
sums			E X = 0.8	$E X^2$ = 0.8

a. Give the formula for E X for a r.v. taking values 0 or 1 only, in terms of p = P(X = 1).
E X = p for a r.v. taking only values 0, 1 with respective probabilities p, q =1 - p.
So E X = 0.8 (confirmed above).

b. Calculate E X = "sum of x p(x)" confirming (a).
Done above.

c. Give the formulas for Var X and sd X (for a r.v. taking values 0 or 1 only) in terms of p = P(X = 1) and q = P(X = 0).
VAr X = pq = 0.8 0.2 = 0.16. Sd X = root(pq) = root(0.16) = 0.4.

d. Calculate Var X and sd X from values of {x, p(x)} confirming your answer (c).
E X = 0.8 = E X^2 as shown in the table above. So Var X = E X^2 - $(E X)^2$ = 0.8 – 0.64 = 0.16. So sd X = root(0.16) = 0.4. This confirms (c).

7.	x	p(x)	x p(x)	x^2 p(x)	$(x - E X)^2$ p(x)
	3	0.4	3 0.4 = 1.2	9 0.4 = 3.6	$(3– 1.0)^2$ 0.4 = 1.6
	-1	0.2	-1 0.2 = –0.2	1 0.2 = 0.2	$(–1– 1.0)^2$ 0.2 = 0.8
	0	0.4	0 0.4 = 0.0	0 0.4 = 0.0	$(0– 1.0)^2$ 0.4 = 0.4
sums			E X = 1.0	E X^2 = 3.8	Var X = 2.8

a. Calculate E X = 1.0 above.

b. Calculate Var X and sd X using the definition (as opposed to the computing formula).
Var X = 2.8 as in the last column above. Sd X = root(Var X) = root(2.8).

c. Re-calculate Var X and sd X using the **computing formula**, confirming (b).
Var X = expected square of X – square of expected X = 3.8 – 1^2 = 2.8 from above, confirming (b).

d. R.v. Y = 3 X – 2. Determine E Y, Var Y and sd Y from your answers above, exploiting known connections with E X, Var X and sd X.
E Y = 3 E X – 2 = 3 (1.0) – 2 = 1.0 also.
Var Y = Var(3 X – 2) = Var(3 X) = 9 Var X = 9 (2.8) = 25.2.
Sd X = root(Var X) = root(9 2.8) = 3 root(2.8) = 3 sd(X).

e. Re-calculate E Y, Var Y and sd Y directly from the table below confirming (d).
(p(x) = p(y) since each y corresponds to precisely one x)

y	x	p(x)	y p(y)	y^2 p(y)
7	3	0.4	7 0.4 = 2.8	49 0.4 = 19.6
-5	-1	0.2	-5 0.2 = -1.0	25 0.2 = 5.0
-2	0	0.4	-2 0.4 = -0.8	4 0.4 = 1.6
sums			E Y = 1.0	E Y^2 = 26.2

So E Y = 1.0 and Var Y = E Y^2 – $(E Y)^2$ = 26.2 – 1.0^2 = 25.2, both as in (d).

8. R.v. X and Y have

E X = 6 E Y = -3

Var X = 9 Var Y = 13

a. Determine E(2 X + 3 Y – 4) = 2 E X + 3 E Y – 4 = 2(6) + 3 (-3) - 4 = -1.

b. If X, Y are independent determine Var(2 X + 3 Y – 4) showing how independence is used.

Var(2 X + 3 Y – 4) = Var(2 X + 3 Y) = (if independent) Var(2 X) + Var(3 Y)

= 4 Var X + 9 Var Y = 4 9 + 9 13 = 153.

c. From (b) determine sd (2 X + 3 Y – 4). = root(153)

9. CLT. Each account of a population of business accounts is scored with x = balance due. Suppose the population mean balance is due is E X = $466.48 population sd sigma = $74.88. A with replacement sample of 100 accounts will be selected.

a. Are you able to sketch the population distribution of x?

No. We know only the mean and sd of the population. There is always the possibility of a population "in control" in which case the distribution is normal with mean 466.48 and sd 74.88. However, there are infinitely many other distributions having these very same mean and sd.

b. Determine E xBAR and sd xBAR.

E xBAR = population mean = 466.48. That is, the sample mean, averaged over all of its possibilities according to their probabilities, is exactly the mean of the population from which the samples are being selected. This is true for with-replacement sampling, as here, but is also true for without-replacement sampling.

sd xBAR = sigma / root(n) = 74.88 / root(100) = 7.488.

c. Sketch the approximation of the distribution of xBAR offered by the central limit theorem (CLT) identifying E xBAR and sd xBAR as recognizable elements of your sketch.

Bell curve (normal density) having mean 466.48 and sd 7.488.

d. Repeat (c) if instead the sample is selected without replacement and the population size is N = 4000. Is your sketch much different from (c) in this case?

Sampling without-replacement alters sd xBAR to

7.488 times FPC = 7.488 root((4000-100)/(4000-1)) = 7.3947 (little changed)

10. CLT. A population of accounts has 30% that are overdue. A sample of 400 accounts is to be selected with-replacement.
a. Sketch the approximate distribution of pHAT = fraction of overdue accounts in the sample. Clearly identify E pHAT and sd pHAT as recognizable elements of your sketch and evaluate them numerically.
pHAT = (number of overdue accounts in the sample of 400) / 400
(it is random since the sample is random).
E pHAT = population fraction of overdue accounts = p = 0.3 (we're told 30% are overdue).
sd pHAT = root(pq / 400) = root(.3 .7 / 400) ~ 0.023.
Ans. Bell curve with mean 0.3 and sd 0.023.
The ACTUAL margin of error for pHAT is therefore +/- 1.96 0.023 ~ 0.045.
Had we not been told p = 0.3 we would estimate ME from the sample as
ME ~ +/- 1.96 root(pHAT qHAT / 400) (often referred to as margin of error)
A 95% z-based CI for p would be pHAT +/- 1.96 root(pHAT qHAT / 400).

b. Repeat (a) except assume that the sample is without-replacement and the population size is 3000. Is this sketch much different from (a)?
FPC = root((3000 – 400) / (3000 – 1)) ~ 0.93.
Bell curve with mean p = 0.3 and sd = 0.023 FPC = 0.023 0.93.

11. We average around 6.4 shortages in a week. The Poisson distribution is thought to apply to X = number of shortages in one week.

a. Determine the probability that there are 5 shortages in one week.
$p(x) = e^{-mu}\ mu^x / x! = e^{-6.4}\ 6.4^5 / 5! \sim 0.149$.

b. Out of 52 weeks, assuming this model applies to all weeks, around how many weeks should experience exactly 5 shortages?
52 times 0.149 = 7.73 ~ 8.

c. Sketch the normal approximation of the distribution of X. Clearly identify E X and sd X (which you evaluate numerically) as recognizable elements of your sketch.
E X = 6.4 > 3. Sd of Poisson = root(mean) = 2.53.
Bell curve having mean 6.4 and sd 2.53 (a fairly good approximation for Poisson with mean > 3).

d. Use the z-table and (c) to approximate P(X in range 3, 4, 5, 6) using the continuity correction. Identify the two z-scores you are using.
z = (2.5 – 6.4) / 2.53 = -1.54
(continuity correction dips back to 2.5 to capture more of p(3) under z-curve)
z = (6.5 – 6.4) / 2.53 = 0.04
p(3) + p(4) + p(5) + p(6) ~ P(-1.54 < Z < 0.04) = P(0 < Z < 1.54) + P(0 < Z < 0.04)
0.4382 + 0.016 = 0.4398

The actual values are p(3) + p(4) + p(5) + p(6)
= 0.0725945 + 0.116151 + 0.148674 + 0.158585 = 0.4960.

e. Use closest z-entry to approximate the 75th percentile of X (you first need to find the 75th percentile of z then convert to an x-score).
75th percentile of Z captures 25% between 0 and itself. So we seek z with P(0 < Z < z) = 0.25. This is the reverse table use. We enter 0.25 to the body of the z-table finding z = 0.67 corresponds to the closest entry 0.2486.

12. A with replacement sample of 50 customers, for score x = dollar value returns they have made to us last year, finds sample mean xBAR = $4.89 with sample sd s = $2.10.
a. Estimate the population mean and population sd.
estimate of pop mean is xBAR = 4.89.
estimate of pop sd is s = 2.10

b. Estimate the sd of xBAR.
theory says sd of xBAR is sigma / root(n)
we estimate sd of xBAR by s / root(n) = 2.10 / root(50) = 0.297.

c. Determine a 95% z-based CI for mu = population mean.
usual xBAR +/- 1.96 s / root(n) = 4.89 +/- 1.96 2.10 / root(50) = [4.30791, 5.47209].
P(mu in (random) CI xBAR +/- 1.96 s / root(n)) ~ 0.95.

13. A **normal** population is sampled finding {3.4, 3.8, 4.8}.
a. Calculate xBAR and s from this data.
xBAR = (3.4 + 3.8 + 4.8) / 3 = 4
s = root(($(3.4 – 4)^2$ + $(3.8 – 4)^2$ + $(4.8 – 4)^2$) / (3 – 1))) = root(0.52) = 0.721.

b. Estimate the population mean and sd from this data.
est of mu is xBAR = 4
est of sigma is s = 0.721

c. Estimate sd xBAR from this data.
est of sd of xBAR is s / root(n) = 0.721 / root(3) = 0.416

d. Determine a 90% t-based CI for the population mean mu from this data.
xBAR +/- t s / root(n) = 4 +/- 2.92 0.721 / root(3) = [2.78449, 5.21551].
DF = 3-1 = 2.

e. What is the value of P(mu in 90% t-based CI) = 0.9
(**exact** if we use t and data calculated to infinite precision).

f. To what sample size nFINAL must we continue in order to obtain a 90% t-based hybrid CI of the form xBARfinal +/- 0.2?
formula $n \sim (t \ s / B)^2 = (2.92 \ 0.721 / 0.2)^2 \sim 111$.

Since this n is so large, it might be best to just toss out the original n0 = 3 samples and sample 111 fresh, then use a z-based CI. We'll not pursue any detailed comparison of the two methods except to say it may be best whenever the indicated nFINAL is large.

g. If we do continue sampling to the nFINAL of part (f) finding xBARfinal = 4.22 what will be our hybrid t-based CI from (f)?
xBARfinal +/- 0.2 = 4.22 +/- 0.2 (really, it is almost exactly this)

14. We desire a 95% z-based CI for p = fraction of viewers of our advertising who have a "favorable impression of our company." A with-replacement sample of 100 viewers finds that 61 have a favorable impression.
a. Give the 95% z-based CI for p based on the above data.
pHAT +/- 1.96 root(pHAT qHAT / n) = 0.61 +/- 1.96 root(0..61 0.39 / 100)
= [0.514401, 0.705599].

b. Determine an nFINAL to which we must continue sampling in order to achieve a 95% z-based CI of the form pHAT +/- 0.1.
$n \sim (z \text{ root}(pHAT \ qHAT) / B)^2 = (1.96 \text{ root}(0.61 \ 0.39) / 0.1)^2 \sim 92$

c. If we do continue sampling to nFINAL finding 63.7% of them have a favorable impression of our company give the resulting 95% hybrid z-based CI for p.

Just quote the CI (a) for the data you have, we have the desired precision, and even a little more, already. Had nFINAL exceeded 100 we'd have used hybrid CI pHATfinal +/- 0.1

15. It is desired to estimate the population average amount our typical corporate business client will spend with us next quarter. We sample100 accounts with-replacement and (carefully) spend some time and money evaluating their likely purchase needs y next quarter. For each account we also have the score x = amount they spent this quarter (on file). From the 100 we find

xBAR = $342000	yBAR = $287000
s(x) = $88000	s(y) = $72000

sample correlation rhoHAT = 0.85
Suppose it is known that mux = 338799 (i.e. avg of all accounts this quarter)
a. Give a 95% z-based CI for mu(y) based on the y-data alone.
yBAR +/- 1.96 sy / root(n) = 287000 +/- 1.96 72000 / root(100) = [272888, 301112].

b. Give the regression estimate of mu(y). Show that it differs from yBAR used in (a).
yBAR + (mux – xBAR) (sy / sx) rhoHAT
= 287000 + (338799 – 342000) (72000 / 88000) 0.85 = 284774 (less than yBAR)
Rationale: Since xBAR has over-estimated mux (known), and x is positively correlated with y, we reduce yBAR.

c. Give the estimate of sd (b) showing that it differs from the estimate of sd(yBAR).
The estimated sd of the regression-based estimator (b) of muy is
(sy / root(n)) root(1 – rhoHAT^2) = (72000 / root(100)) root(1 – 0..85^2) = 3793.
This is less than the estimated sd of yBAR which is 7200. It is as though instead of 100 samples we'd had 100 (7200 / 3793)^2 ~ 360. That is a big advantage held by the regression method. However, you'd need to know mux and sample x-scores.

d. Give the 95% z-based regression-estimator based CI for mu(y) showing that it is narrower than (a) for the same sample size 100.
regr est +/- 1.96 (sy / root(n)) root(1 – rhoHAT^2) = 284774 +/- 3793 = [280981, 288567].
This CI is far narrower than the usual CI (a) based upon yBAR alone. To use the regression based approach we have to know mux and be able to pick up x-scores along with y scores, but the savings can be considerable if rhoHAT is near –1 or 1.

e. Modify (d) if the sample is without-replacement and the population size is 3000.
regr est +/- 1.96 (sy / root(n)) root(1 – rhoHAT^2) FPC
= 284774 +/- 1.96 3793 root((3000-100) / (3000-1)) = [281044, 288504]
It has FPC = 0.983356 times width of the regr-based CI for with-replacement sampling.

16. **Independent** samples of 60 women and 40 men, all professionals living alone, are drawn with-replacement finding

women	xBAR = 24	s(x) = 15	n(x) = 60
men	yBAR = 18	s(y) = 10	n(y) = 40

where x = amount spent dining out, y = amount spent dining out.
a. Give an estimate of mu(x) – mu(y).
A usual estimator is xBAR-yBAR = 24 – 18 = 6.
In place of xBAR we sometimes use more sophisticated estimators, not discussed here, such a "toss out the very largest and smallest of the data then average the rest." Such alternative estimators are used to guard against "outliers" (e.g. Bill Gates throws off an average income if he is included but the estimator just described would leave him out).

b. Give estimates of (population) sigmaWOMEN and sigmaMEN.
est of sigmaWOMEN is sWOMEN = 15
est of sigmaMEN is sMEN = 10.

c. Give estimates of sd xBAR, sd yBAR, sd(xBAR – yBAR).
est of sd of xBAR is s(x)/root(nx) = 15 / root(60) = 1.94
est of sd of yBAR is s(y) / root(ny) = 10 / root(40) = 1.58
est of sd of (xBAR – yBAR) is root(s(x)^2 / n(x) + s(y)^2 / n(y))
= root(225 / 60 + 100 / 40) = 2.5 (more variable than each of xBAR, yBAR alone)

d. Give 95% z-based CI for mu(x)-mu(y).
(xBAR – yBAR) +/- 1.96 est sd of (xBAR – yBAR) = 6 +/- 1.96 2.5 = [1.1, 10.9]
which is uncomfortably wide compared with the estimate xBAR – yBAR = 6.

17. **Independent** samples of 60 women and 40 men, all working the same job, are drawn with-replacement finding

women 35 have hand cramping
men 17 have hand cramping

Define p(x) = fraction of all women in this job who have hand cramping, p(y) the corresponding rate for men.
a. Estimate each of p(x), p(y).
pHATx = 35 / 60 = 0.58333
pHATy = 17 / 40 = 0.425

b. Give estimates of sd pHATx , sd pHATy and pHATx-pHATy.
est of sd of pHATx is root(pHATx qHATx / nx) = root(.58333 .41667 / 60) = 0.064
est of sd of pHATy is root(pHATy qHATy / ny) = root(.425 .575 / 40) = 0.078
est of sd of (pHATx – pHATy) is root(pHATx qHATx / nx + pHATy qHATy /ny)
= root(.58333 .41667 / 60 + .425 .575 / 40) = 0.101 (larger than for either pHAT)

c. Give a 90% z-based CI for p(x)-p(y).
(pHATx – pHATy) +/- 1.645 est sd of (pHATx – pHATy)
(.58333 - .425) +/ - 1.645 0.101 = [-0.007815, 0.324475].