UNIVERSITY OF TORONTO
FACULTY OF APPLIED SCIENCE AND ENGINEERING

FINAL EXAMINATION, April 2017

DURATION: 2 Hours 30 Minutes

Second Year — Engineering Science

STA286H1 S — Probability and Statistics

Calculator Type: 2
Exam Type: B

Examiner: N. Montgomery

1. Please write your name EXACTLY as it appears on your student card.
2. There are 15 pages including this page.
3. Feel free to use the backs of the question pages for rough work.
4. Don't ask me how small a p-value has to be.
5. There are 60 marks in total. I have indicated the number of marks available for each part of each question.

FAMILY NAME:_____

GIVEN NAME:_____

STUDENT NUMBER:_____

Q.	Marks
1.	
2.	
3.	
4.	
5.	
Total	

1. **(10 marks total)** Suppose S is some sample space and P is a probability function. Suppose $P(A) = 0.3$ and $P(A|B) = 0.7$.

(a) **(4 marks)** If there is a largest possible value for $P(B)$, determine what it is. If not, say why not.

(b) **(2 marks)** If there is a smallest possible value for $P(B)$, determine what it is. If not, say why not.

(c) **(4 marks)** Suppose we find out that $P(B) = 0.4$. Calculate $P(B^c|A)$.

2. **(10 marks total)** One of the earliest uses of Poisson process was in the study of radioactive decay. One gram of the isotope Americium-241 undergoes spontaneous fission according to a Poisson process at a rate of 1.2 decays per second. Spontaneous fission can be detected using a "scintillation detector". Or at least this is what the internet says.

   Anyway, your detector keeps a database of the number of fission events and also the times at which they occur. The database will also keep track of equipment malfunctions and other hings that might happen.

   You turn on your detector and start to detect fission events.

(a) **(2 marks)** What is the probability that any particular 2 second interval will contain 0 fission events?

(b) **(3 marks)** A fission event occurs at 318.6 seconds. The next fission event occurs at some time before 322.1 seconds; however, the timer had malfunctioned at 319.4 seconds, so you cannot be sure when exactly it occurred. What is expected time at which the fission event occurred

(c) **(3 marks)** What is the probability that 3 or more fission events happen between 400 and 406 seconds?

(d) **(2 marks)** Suppose 80 fission events happen between 500 and 600 seconds. Using your intuition (without proving anything), what do you think is the distribution of the number of events that occurred between 500 and 520 seconds, given 80 occurred between 500 and 600 seconds.

3. **(10 marks total)** The confidence intervals we saw in this course were all of the form

"estimate" $\pm$ "margin of error"

When the sample is from a $N(\mu, \sigma)$ population, the sample size required to get a 95% C.I. with an absolute margin of error e is given by:

$$n = \left(\frac{z_{0.025}\sigma}{e}\right)^2$$

There was then the problem of the unknown σ, etc. In this question we're going to deal with the case of estimating a probability p from a Bernoulli(p) population rather than the mean from a normal population.

(a) **(2 marks)** Show that to produce a 95% confidence interval for p with absolute margin of error e, the sample size required is:

$$n = \left(\frac{z_{0.025}\sqrt{p(1-p)}}{e}\right)^2$$

(b) **(2 marks)** The formula in (a) has a familiar problem. It depends on an unknown parameter. There is a common procedure to handle this problem in practice, as follows:

> Determine, from your knowledge of the underlying problem, a range of plausible values that p might take on. Say the range is $R = [p_1, p_2]$. If $0.5 \in R$, plug $p = 0.5$ into the sample size formula. If $0.5 \notin R$, plug in whichever out of p_1 and p_2 is closer to 0.5.

For example, suppose you think a range of plausible values is $R = [0.2, 0.6]$. Since $0.5 \in R$, you would use $p = 0.5$ in the sample size formula. But if your range of plausible values were $R = [0.6, 0.9]$, you would use $p = 0.6$ in the sample size formula.

Use the sample size formula to justify the use of this procedure.

(c) **(2 marks)** A gas company wants to estimate the probability that one of its gas meters has any rust on the cover. This is a very common fault. The gas company thinks the probability is somewhere between 0.2 and 0.4. Determine the sample size required to produce a 95% confidence interval for p with a margin of error of 0.05, using the formula and procedure from (b) and (c).

If you are unable to calculate a sample size, just use $n = 80$ in the following question.

(d) **(2 marks)** The gas company gathers a sample of the size you specified in (c) and found 23 gas meters with rusty covers. Produce a 95% confidence interval for the probability that a gas meter cover has rust.

(e) **(2 marks)** It is very unlikely that a gas meter is not properly attached to the building it belongs too. That would be a serious installation fault. The gas company thinks the probability of such a meter is between 0 and 0.03. They want to estimate the true probability with a 95% confidence interval with the same margin of error as used in (c). Explain why the formula and procedure from (b) and (c) give a sample size that probably isn't useful in this case.

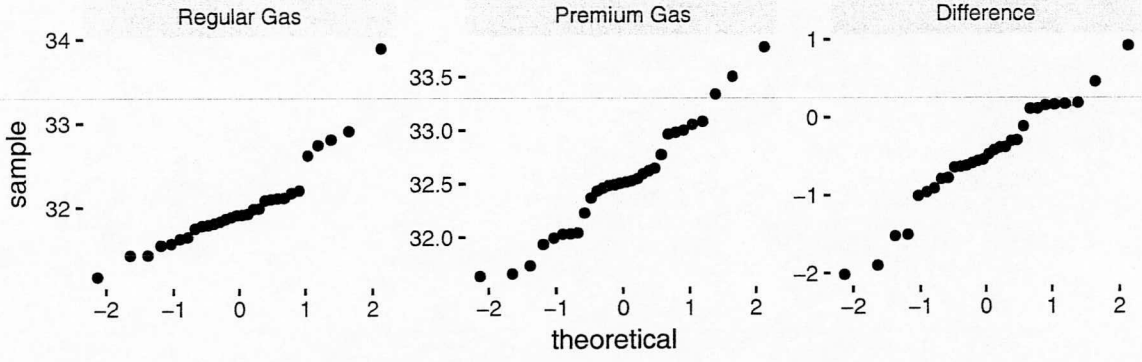
4. **(10 marks total)** A gasoline manufacturer hires a ***summer student*** (!) to manage the data collection and analysis for an experiment it is conducting to see if there is a difference in fuel economy when cars use regular gasoline or premium gasoline.

60 similar Toyota Corollas (a brand of small passenger car) are divided into two groups: `Regular Gas` and `Premium Gas`. Every day a car from the `Regular Gas` group gets regular gasoline and drives around a track at for 400 km at 80 km/h. Then, a car from the `Premium Gas` group does the same thing. The amount in liters of fuel used by each car is recorded in the dataset. After 30 days the experiment is done.

The student uses a mass market spreadsheet to record the data and to do some analysis. Here is the structure of the spreadsheet the student puts together:

Day	`Regular Gas`	`Premium Gas`	Difference
1	33.37	32.86	0.52
2	31.44	33.10	-1.67
⋮	⋮	⋮	⋮
30	31.36	32.68	-1.32
Sample Average	32.02	32.527	-0.507
Sample Standard Deviation	0.534	0.527	0.666

Here are normal quantile plots for the `Regular Gas`, `Premium Gas`, `Difference` Columns.



(a) **(3 marks)** Do you think this experiment has one or two independent samples? Briefly justify your answer. (I think there is a good argument either way. **Choose one**. Justify it. Briefly. Move on.)

(b) **(5 marks)** Use you choice in (a) to perform the hypothesis test with the null hypothesis corresponding to "no difference between the means" of the two groups, using a p-value in your conclusion.

(c) **(2 marks)** What assumptions are required for your probability calculation in (b) to be accurate, and to what extent can you say they are satisfied in this case?

5. **(20 marks total)** Let X be a random variable with a mean $\mathrm{E}(X)$ and a variance $\mathrm{Var}(X)$.

(a) **(3 marks)** The definition of variance is $\mathrm{Var}(X) = \mathrm{E}\left((X - \mathrm{E}(X))^2\right)$. Prove, using only this definition, some basic algebra, and properties of expected values, that $\mathrm{E}\left(X^2\right) = \mathrm{Var}(X) + (\mathrm{E}(X))^2$.

(b) **(2 marks)** Suppose further that X has a moment generating function $M_X(t) = \mathrm{E}\left(e^{tX}\right)$. Prove that the moment generating function for aX with $a \neq 0$ is $M_{aX}(t) = M_X(at)$. (This has nothing to do with (a).)

(c) **(3 marks)** Suppose even further that $X \sim \text{Gamma}(\alpha, \lambda)$ is the underlying distribution for an independent and identically distributed (i.i.d.) sample $X_1, \ldots, X_n$. Prove, for any constant $c > 0$, that:

$$\frac{\sum_{i=1}^{n} X_i}{c} \sim \text{Gamma}(n\alpha, c\lambda)$$

(Use what you know about moment generating functions, along with the result from (b))

(d) **(6 marks)** Suppose even *further* that we know $\alpha = 2$. We want to estimate the parameter $\theta = 1/\lambda^2$. We observe data $x_1, \ldots, x_n$. Show that the maximum likelihood estimator (MLE) for θ is:

$$\hat{\theta} = \left(\frac{\sum_{i-1}^{n} X_i}{2n} \right)^2$$

(I suggest you find the MLE for λ and use the invariance property that MLEs enjoy.)

(e) **(4 marks)** Determine $E(\hat{\theta})$. (Now is the time to use the result from (a), along with the result from (c).)

(f) **(2 marks)** What is the unbiased estimator for θ with the smallest variance?

In all that follows I might use $q = 1 - p$ when I feel like it.

name	pmf/pdf	support	mean	variance	mgf $M(t)$ or $M(s)$
Bernoulli(p)	$p^x(1-p)^{1-x}$	$x \in \{0,1\}$	p	$p(1-p)$	$q + pe^t$
Binomial(n,p)	$\binom{n}{k}p^k(1-p)^{n-k}$	$k \in \{0,\ldots,n\}$	np	$np(1-p)$	$(q+pe^t)^n$
Geometric(p)	$(1-p)^{k-1}p$	$k \in \{1,2,\ldots\}$	$1/p$	q/p^2	$\frac{pe^t}{1-qe^t}$
NegBin(r,p)	$\binom{k-1}{r-1}(1-p)^{k-r}p^r$	$k \in \{r, r+1, r+2, \ldots\}$	r/p	rq/p^2	$\left(\frac{pe^t}{1-qe^t}\right)^r$
Poisson Process $N(t)$	$\frac{(\lambda t)^k}{k!}e^{-\lambda t}$	$k \in \{0,1,2,\ldots\}$	λ	λ	$M(s) = e^{\lambda t(e^s-1)}$
Uniform$[a,b]$	$\frac{1}{b-a}$	$a < x < b$	$\frac{a+b}{2}$	$\frac{(b-a)^2}{12}$	not needed
Exp(λ)	$\lambda e^{-\lambda x}$	$x > 0$	$1/\lambda$	$1/\lambda^2$	$\frac{\lambda}{\lambda - t}$
Gamma(α,λ)	$\frac{\lambda^\alpha}{\Gamma(\alpha)}x^{\alpha-1}e^{-\lambda x}$	$x > 0$	α/λ	α/λ^2	$\left(\frac{\lambda}{\lambda-t}\right)^\alpha$
N(μ,σ)	$\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$	$x \in \mathbb{R}$	μ	σ^2	$e^{\mu t + \sigma^2 t^2/2}$

Note that the textbook uses $\beta = 1/\lambda$ when dealing with Exponential and Gamma distributions.

Here is a table of the common "statistics" we used:

Statistic	(Approx.) Dist^n	Comment	Use
$\frac{\overline{X}-\mu}{s/\sqrt{n}}$	t_{n-1}	Sample size formula: $(z_{\alpha/2}\sigma/e)^2$	Single normal sample
$\frac{\hat{p}-p}{\sqrt{p(1-p)/n}}$	$N(0,1)$		Single Bernoulli sample
$\frac{(\overline{X}_1-\overline{X}_2)-(\mu_1-\mu_2)}{s_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}$	$t_{n_1+n_2-2}$	$s_p^2 = \frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{n_1+n_2-2}$	Two normal samples
$\frac{(\overline{X}_1-\overline{X}_2)-(\mu_1-\mu_2)}{\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}}$	t_ν	ν has a nasty formula	Two normal samples
$\frac{(\hat{p}_1-\hat{p}_2)-(p_1-p_2)}{\sqrt{p_1(1-p_1)/n_1+p_2(1-p_2)/n_2}}$	$N(0,1)$	"Pooled" $\hat{p} = \frac{x_1+x_2}{n_1+n_2}$	Two Bernoulli samples

t probabilities

	Upper tail probabilities for t_ν distributions $P(t_\nu \geq t)$										
df	0.2	0.15	0.1	0.05	0.025	0.0125	0.01	0.0075	0.005	0.001	0.0005
10	0.879	1.093	1.372	1.812	2.228	2.634	2.764	2.932	3.169	4.144	4.587
11	0.876	1.088	1.363	1.796	2.201	2.593	2.718	2.879	3.106	4.025	4.437
12	0.873	1.083	1.356	1.782	2.179	2.560	2.681	2.836	3.055	3.930	4.318
13	0.870	1.079	1.350	1.771	2.160	2.533	2.650	2.801	3.012	3.852	4.221
14	0.868	1.076	1.345	1.761	2.145	2.510	2.624	2.771	2.977	3.787	4.140
15	0.866	1.074	1.341	1.753	2.131	2.490	2.602	2.746	2.947	3.733	4.073
16	0.865	1.071	1.337	1.746	2.120	2.473	2.583	2.724	2.921	3.686	4.015
17	0.863	1.069	1.333	1.740	2.110	2.458	2.567	2.706	2.898	3.646	3.965
18	0.862	1.067	1.330	1.734	2.101	2.445	2.552	2.689	2.878	3.610	3.922
19	0.861	1.066	1.328	1.729	2.093	2.433	2.539	2.674	2.861	3.579	3.883
20	0.860	1.064	1.325	1.725	2.086	2.423	2.528	2.661	2.845	3.552	3.850
21	0.859	1.063	1.323	1.721	2.080	2.414	2.518	2.649	2.831	3.527	3.819
22	0.858	1.061	1.321	1.717	2.074	2.405	2.508	2.639	2.819	3.505	3.792
23	0.858	1.060	1.319	1.714	2.069	2.398	2.500	2.629	2.807	3.485	3.768
24	0.857	1.059	1.318	1.711	2.064	2.391	2.492	2.620	2.797	3.467	3.745
25	0.856	1.058	1.316	1.708	2.060	2.385	2.485	2.612	2.787	3.450	3.725
26	0.856	1.058	1.315	1.706	2.056	2.379	2.479	2.605	2.779	3.435	3.707
27	0.855	1.057	1.314	1.703	2.052	2.373	2.473	2.598	2.771	3.421	3.690
28	0.855	1.056	1.313	1.701	2.048	2.368	2.467	2.592	2.763	3.408	3.674
29	0.854	1.055	1.311	1.699	2.045	2.364	2.462	2.586	2.756	3.396	3.659
30	0.854	1.055	1.310	1.697	2.042	2.360	2.457	2.581	2.750	3.385	3.646
31	0.853	1.054	1.309	1.696	2.040	2.356	2.453	2.576	2.744	3.375	3.633
32	0.853	1.054	1.309	1.694	2.037	2.352	2.449	2.571	2.738	3.365	3.622
33	0.853	1.053	1.308	1.692	2.035	2.348	2.445	2.566	2.733	3.356	3.611
34	0.852	1.052	1.307	1.691	2.032	2.345	2.441	2.562	2.728	3.348	3.601
35	0.852	1.052	1.306	1.690	2.030	2.342	2.438	2.558	2.724	3.340	3.591
36	0.852	1.052	1.306	1.688	2.028	2.339	2.434	2.555	2.719	3.333	3.582
37	0.851	1.051	1.305	1.687	2.026	2.336	2.431	2.551	2.715	3.326	3.574
38	0.851	1.051	1.304	1.686	2.024	2.334	2.429	2.548	2.712	3.319	3.566
39	0.851	1.050	1.304	1.685	2.023	2.331	2.426	2.545	2.708	3.313	3.558
40	0.851	1.050	1.303	1.684	2.021	2.329	2.423	2.542	2.704	3.307	3.551
41	0.850	1.050	1.303	1.683	2.020	2.327	2.421	2.539	2.701	3.301	3.544
42	0.850	1.049	1.302	1.682	2.018	2.325	2.418	2.537	2.698	3.296	3.538
43	0.850	1.049	1.302	1.681	2.017	2.323	2.416	2.534	2.695	3.291	3.532
44	0.850	1.049	1.301	1.680	2.015	2.321	2.414	2.532	2.692	3.286	3.526
45	0.850	1.049	1.301	1.679	2.014	2.319	2.412	2.529	2.690	3.281	3.520
46	0.850	1.048	1.300	1.679	2.013	2.317	2.410	2.527	2.687	3.277	3.515
47	0.849	1.048	1.300	1.678	2.012	2.315	2.408	2.525	2.685	3.273	3.510
48	0.849	1.048	1.299	1.677	2.011	2.314	2.407	2.523	2.682	3.269	3.505
49	0.849	1.048	1.299	1.677	2.010	2.312	2.405	2.521	2.680	3.265	3.500
50	0.849	1.047	1.299	1.676	2.009	2.311	2.403	2.519	2.678	3.261	3.496
51	0.849	1.047	1.298	1.675	2.008	2.310	2.402	2.518	2.676	3.258	3.492
52	0.849	1.047	1.298	1.675	2.007	2.308	2.400	2.516	2.674	3.255	3.488
53	0.848	1.047	1.298	1.674	2.006	2.307	2.399	2.514	2.672	3.251	3.484
54	0.848	1.046	1.297	1.674	2.005	2.306	2.397	2.513	2.670	3.248	3.480
55	0.848	1.046	1.297	1.673	2.004	2.304	2.396	2.511	2.668	3.245	3.476
56	0.848	1.046	1.297	1.673	2.003	2.303	2.395	2.510	2.667	3.242	3.473
57	0.848	1.046	1.297	1.672	2.002	2.302	2.394	2.508	2.665	3.239	3.470
58	0.848	1.046	1.296	1.672	2.002	2.301	2.392	2.507	2.663	3.237	3.466
59	0.848	1.046	1.296	1.671	2.001	2.300	2.391	2.506	2.662	3.234	3.463
60	0.848	1.045	1.296	1.671	2.000	2.299	2.390	2.504	2.660	3.232	3.460
∞	0.842	1.036	1.282	1.645	1.960	2.241	2.326	2.432	2.576	3.090	3.291