Princeton University
Department of Operations Research
and Financial Engineering

ORF 245
Fundamentals of Statistics

Practice Final Exam

January ??, 2016
1:00 – 4:00 pm

Closed book. No computers. Calculators allowed.

You are permitted to use a two-page two-sided cheat sheet.

Return the exam questions and your cheat sheet with your exam booklet.

ADVICE: Spend five minutes now to read the exam from beginning to end. It only takes a few minutes and will give you a good idea of the scope of the exam and that will help you pace yourself through it.

NOTE: I have tried to make the problems "interesting". Interesting problems are often more difficult than "routine" problems.

---

(1) *(10 pts.)* Who's your mamma?

(2) *(5 pts.)* Suppose that X, Y, and Z are normal random variables with these means and variances:

	mean	variance
X	2	4
Y	0	9
Z	15	16

Rank the following probabilities (smallest to largest):

$$P(X \leq 1), \quad P(Y \leq -2), \quad P(Z \leq 12)$$

(3) *(10 pts.) Twin studies:* Many characteristics of individuals are determined by genetics, but many others are affected by their environment. There is, therefore, much interest in comparing identical twins who have been raised apart.

The table below shows the IQs of ten pairs of identical twins who were raised apart. In each pair, one twin had been raised in a "good" environment and another in a "poor" environment.

	IQ	
Family	Poor environment	Good environment
1	100	125
2	65	95
3	60	100
4	125	120
5	85	120
6	145	185
7	55	80
8	180	210
9	60	105
10	135	175

The genetic influence on IQ is evident – when one twin has high IQ, the other often does too. However we can also ask...

Do the twins raised in a "good" environment have a different mean IQ from those raised in a "poor" environment?

To answer this question, let X denote the difference in IQ of the "good" environment minus the IQ of the twin from the "bad" environment. Test the hypothesis that the mean μ_X is zero.

(4) *(5 pts.)* Recall that a random variable Y has a log-normal distribution if $Y = e^{\mu+\sigma X}$ where X is a standard normal variable (i.e, X has mean zero and variance one). Compute the formula for the pdf of a log-normal distribution.

(5) *(10 pts.)* In class we discussed confidence intervals for the parameter p of a Bernoulli random variable. The straight-forward approach relies on the fact that the mean μ equals the parameter p and therefore the usual technique of appoximating the variance σ^2 by the sample variance s^2 gives the standard confidence interval. But, we also introduced what we called a "better" confidence interval, which is based on the fact that the variance also has a simple formula relating it to the parameter p and this was used to avoid the approximation of σ^2. The same trick can be applied to other situations in which the underlying distribution has just a single parameter so that both the mean and the variance can be expressed in terms of this parameter. The Poisson distribution is such an example. Let X be a Poisson random variable with unknown parameter λ. Recall that $\mathbb{E}(X) = \lambda$ and $\mathbb{V}\text{ar}(X) = \lambda$. Recall further that, if n is large, then the central limit theorem can be invoked to get the standard

$100(1-\alpha)\%$-confidence interval for λ:

$$\overline{X} - z_{\alpha/2} S/\sqrt{n} \;\leq\; \lambda \;\leq\; \overline{X} + z_{\alpha/2} S/\sqrt{n}.$$

Of course, we could use the fact that $\mathbb{V}\text{ar}(X) = \lambda = \mathbb{E}(X)$ to replace S with $\sqrt{\overline{X}}$:

$$\overline{X} - z_{\alpha/2} \sqrt{\overline{X}/n} \;\leq\; \lambda \;\leq\; \overline{X} + z_{\alpha/2} \sqrt{\overline{X}/n}.$$

(a) Under the same assumption that n is large, derive a "better" confidence interval for λ.
(b) Comment on how similar or different the better interval is relative to the standard interval.

(6) *(10 pts.)* Suppose you are given a coin and told that it has a 2-to-1 bias. That is, the coin favors one side coming up that way 66.66% of the time. But you weren't told if it favors heads or it favors tails. You must decide. Of course, you could toss the coin a zillion times and it would be obvious which side it favors. But, to make a simplified exam question, let's assume you decide to flip the coin $n = 7$ times and make a decision based on the outcome of these seven flips.

(a) Formulate a null hypothesis, H_0 and an alternative hypothesis, H_1.
(b) Let X denote the number of times the coin comes up heads in the $n = 7$ flips. Assuming that Type-I and Type-II errors are equally bad, formulate a reasonable H_0 rejection region based on the test statistic X.
(c) What is the probability of a Type-I error?
(d) What is the probability of a Type-II error?

NOTE: Here is the probability mass function for a Binomial random variable with $n = 7$ and $p = 2/3$.

x	0	1	2	3	4	5	6	7
p(x)	0.0005	0.0064	0.0384	0.1280	0.2561	0.3073	0.2048	0.0585

(7) *(5 pts.)* Consider the same coin described in the previous problem. Now suppose that the coin is tossed 100 times and comes up heads 60 times. What's a 95% confidence interval for the probability p that the coin comes up heads?

(8) *(15 pts.)* Suppose that $X_1, X_2, \ldots, X_n$ are independent identically distributed with density function

$$f(x) = \begin{cases} \frac{1}{\beta} e^{-x/\beta}, & x \geq 0 \\ 0, & x < 0. \end{cases}$$

(a) Compute a formula for the mean $\mu = \mathbb{E}(X_i)$ of the X_i's.

(b) Compute a formula for the variance $\sigma^2 = \mathbb{E}((X_i - \mu)^2)$ of the X_i's.
(c) What would you suggest as a good estimator for β?
(d) Here is a sample of size $n = 49$:

0.0496	0.1271	0.2060	0.3845	0.1117	0.3066	0.1802
0.0379	0.0544	0.0340	0.0200	0.0649	2.1014	1.8396
0.7756	0.4279	1.2035	0.2731	0.3424	0.1953	0.0244
0.3848	0.3452	0.0482	0.3429	0.6301	0.2157	0.0492
0.3044	0.1155	0.1481	0.2539	0.2379	0.0164	0.7903
0.4054	0.2563	0.1693	0.0415	0.2409	0.2632	0.0932
0.2417	0.0477	0.2427	0.2218	0.2287	0.4685	0.7826

For your convenience, here is the first and second moments of these 49 numbers:

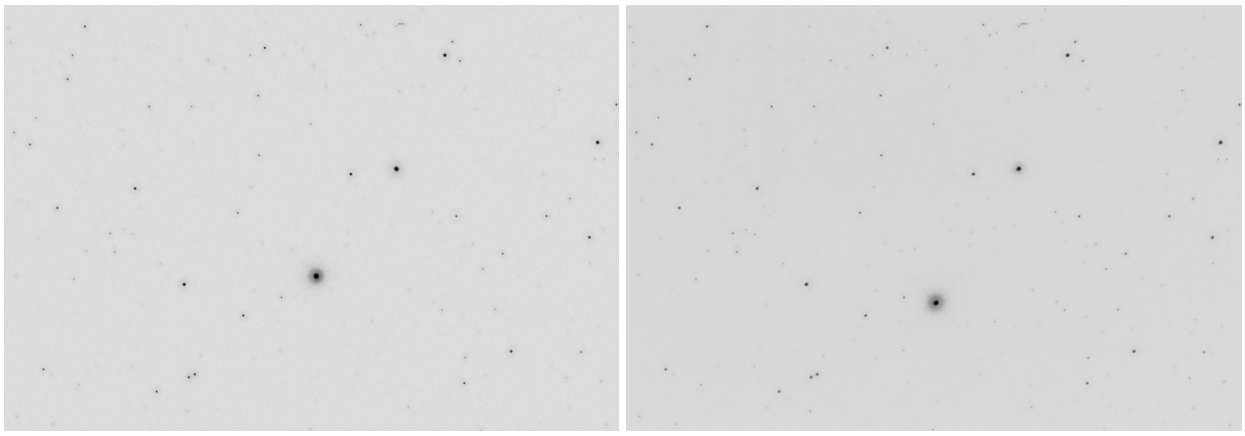
$$\overline{x} = 0.3336 \qquad \overline{x^2} = 0.2812$$

Find a 95% confidence interval for β.

Hint: In case you have forgotten how to integrate simple exponential functions, here's a few precomputed integrals you might need:

$$\int_0^\infty e^{-x}dx = 1, \quad \int_0^\infty xe^{-x}dx = 1, \quad \int_0^\infty x^2e^{-x}dx = 2, \quad \int_0^\infty x^3e^{-x}dx = 6.$$

(9) *(15 pts.)* The stars we observe in the night sky orbit around our Milky Way galaxy once every few hundred million years. If all the stars circled the Milky Way in lock step, then we wouldn't notice any apparent variations in the night sky (ignoring, of course, the daily once-around rotation caused by our own Earth's rotation). But they don't. There's a certain amount of randomness to the whole process. The stars that are far from us don't appear to us to be moving (at least not on human time scales) but some of the nearby stars exhibit rather significant apparent motions relative to the background stars. The star with the greatest apparent motion is Barnard's star. Here are two pictures of it that I took through a 10" telescope on my driveway about two years apart:



Barnard's star is the brightest star in these two pictures (shown as black since I'm showing the "negative" of the true image). Notice how all the other stars appear not to have moved (come back in a few hundred thousand years and they all will show some motion). Only Barnard's star has moved—downward. We can, and I have, use the background stars to set a frame of reference. Actually, I didn't just take two pictures—I took six pictures over the course of a little more than two years. Here's a table showing the time at which each picture was taken (in years) and the x and y coordinates, in pixels, of Barnard's star:

t	x	y
-0.494	681.527	-639.539
0.461	680.374	-656.999
0.715	678.639	-661.234
1.307	679.479	-672.151
1.547	677.950	-676.775
1.844	676.727	-681.713

We can make two regression models: one for x as a linear function of t and the other for y as a linear function of t. For example, here's the regression model for x:

$$X_i = \alpha_x + \beta_x t_i + \epsilon_i$$

(a) Using the data given above, compute the estimators $\widehat{\alpha}_x$ and $\widehat{\beta}_x$ of α_x and β_x.

(b) Give a 95% confidence interval for β_x.

(c) Similarly, derive a 95% confidence interval for the corresponding slope coefficient β_y in the y-model.

(d) From the focal-length of the telescope and the camera's pixel size, one can convert the numbers computed above to standard units of *arcseconds/year* for the so-called *proper motion* of Barnard's star. The conversion formula for the particular telescope/camera used is

$$\text{Proper Motion} = \sqrt{\beta_x^2 + \beta_y^2} \times 0.575$$

Compute the proper motion in arcseconds-per-year.

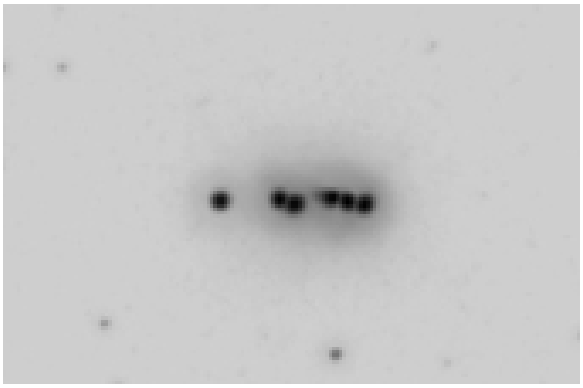
(Note: Four out of four AI's surveyed did not know what an arcsecond is. An arcsecond is a unit of angular measurement. It is a small fraction of a circle. There are 360 degrees in a full circle. There are 60 *arcminutes* in one degree and there are 60 *arcseconds* in one arcminute. So, there are 3600 arcseconds in one degree. As a point of reference, the Moon's apparent diameter in the sky is about $1/2$ degree or, in other words, about 1800 arcseconds.

(e) Explain how you would use bootstrap to produce a 95% confidence interval for the proper motion. (Note: Pseudo-code in your language of choice is encouraged here.)

PS. For your computational convenience, we have precomputed some of things you might need:

$$\begin{array}{rclrclrcl} \overline{t} & = & 0.8967, & \overline{X} & = & 679.1160, & \overline{Y} & = & -664.7352 \\ \overline{t^2} & = & 1.4116, & \overline{X^2} & = & 461201, & \overline{Y^2} & = & 442072 \\ \overline{tX} & = & 607.8262, & \overline{tY} & = & -607.0463, & \overline{XY} & = & -451410 \end{array}$$

(10) *(15 pts.)* This problem is a continuation of the previous one. Here's a highly zoomed-in picture showing all six observations overlayed onto one picture and rotated so that the apparent proper motion is from left to right:

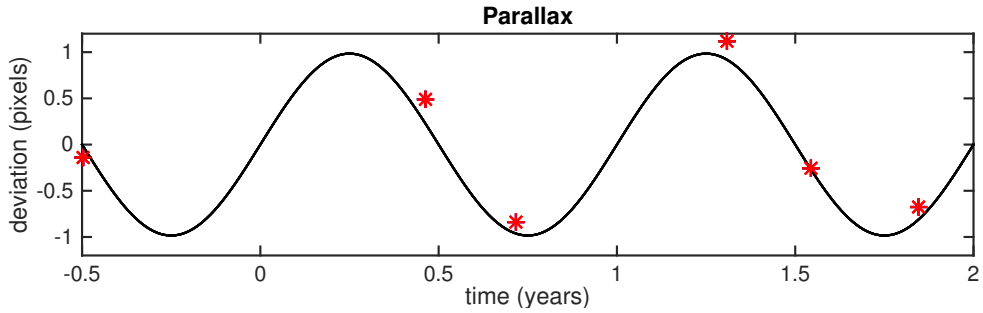


Note that there seems to be a systematic, perhaps sinusoidal, oscillation. This oscillation is no accident. It has a period of exactly one year and it is because the Earth is going around the Sun once a year. Hence, stars that aren't too far away (such as Barnard's star) have an apparent annual wobble as viewed against the more distant background stars (the background stars also have a wobble, but it is tiny and therefore unnoticeable). If we can accurately measure the amplitude of this wobble, we can derive the actual distance to Barnard's star.

The first step to achieving this goal is to subtract the estimated proper motion from the data and then look at the deviation from zero in this adjusted data set. We have done that for you. Here are the six deviations (which, for lack of a better letter, we'll denote by z):

t	z
-0.4940	-0.1410
0.4610	0.4833
0.7150	-0.8375
1.3070	1.1157
1.5470	-0.2668
1.8440	-0.6741

Here's a plot of these six data points together with the best sinusoidal fit to the data:



The goal here is to find a formula for the statistical estimator of the amplitude of the sine wave. The regression model has a simple form:

$$Z_i = \alpha \sin(2\pi t_i) + \varepsilon_i.$$

(a) Derive a least-squares regression formula for an estimator $\widehat{\alpha}$ for the true amplitude α.

(b) Use the data above to compute $\widehat{\alpha}$.

(c) Explain how you would use bootstrap to produce a 95% confidence interval for α. (Note: Pseudo-code in your language of choice is encouraged here.)

(d) One of the standard units for measuring distance in astronomy is the so-called *parsec* (1 parsec $= 3.26$ lightyears). An object is one parsec away if, by definition, the amplitude of its parallax wobble is one arcsecond. It is 10 parsecs away if its parallax is $1/10$-th of an arcsecond—smaller parallax means greater distance. In general, the definition of the distance in parsecs is the reciprocal of the parallax angle expressed in arcseconds. Again, the conversion from pixels to arcseconds requires multipling the number of pixels by 0.575. Hence, the formula for distance in parsecs is:

$$\text{Parallax Distance} = 1/(\alpha \times 0.575).$$

What is your computed distance to Barnard's star in parsecs?
(Comment: In the Star Wars movies, the term *parsec* is used as a unit of time, not a unit of distance. It's one of the things George Lucas got wrong.)

Linear Regression Formulas

Here are some useful formulas you probably have on your cheat sheet...
Consider the regression model:

$$Y_i = \alpha + \beta x_i + \varepsilon_i$$

The least-squares regression formulas for the estimators $\widehat{\alpha}$ and $\widehat{\beta}$ are

$$\begin{aligned} \widehat{\beta} &= \frac{\overline{xY} - \overline{x}\,\overline{Y}}{\overline{x^2} - \overline{x}^2} \\ \widehat{\alpha} &= \overline{Y} - \widehat{\beta}\,\overline{x} \end{aligned}$$

And here are formulas for the associated $100(1-\alpha)\%$-confidence intervals...

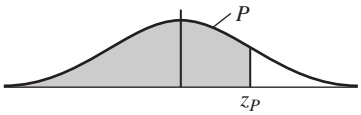
$$\widehat{\alpha} - t_{n-2}(\alpha/2)\,\frac{S}{\sqrt{n}}\,\frac{\sqrt{\overline{x^2}}}{\sqrt{\overline{x^2} - \overline{x}^2}} \;\leq\; \alpha \;\leq\; \widehat{\alpha} + t_{n-2}(\alpha/2)\,\frac{S}{\sqrt{n}}\,\frac{\sqrt{\overline{x^2}}}{\sqrt{\overline{x^2} - \overline{x}^2}}$$

$$\widehat{\beta} - t_{n-2}(\alpha/2)\,\frac{S}{\sqrt{n}}\,\frac{1}{\sqrt{\overline{x^2} - \overline{x}^2}} \;\leq\; \beta \;\leq\; \widehat{\beta} + t_{n-2}(\alpha/2)\,\frac{S}{\sqrt{n}}\,\frac{1}{\sqrt{\overline{x^2} - \overline{x}^2}}$$

where

$$S^2 = \frac{1}{n-2}\sum_i \left(Y_i - \widehat{\alpha} - \widehat{\beta} x_i\right)^2$$

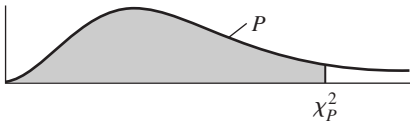
TABLE 2 Cumulative Normal Distribution—Values of P Corresponding to z_p for the Normal Curve



z is the standard normal variable. The value of P for $-z_p$ equals 1 minus the value of P for $+z_p$; for example, the P for –1.62 equals 1 – .9474 = .0526.

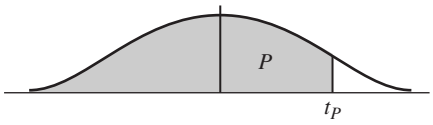
z_p	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
.0	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279	.5319	.5359
.1	.5398	.5438	.5478	.5517	.5557	.5596	.5636	.5675	.5714	.5753
.2	.5793	.5832	.5871	.5910	.5948	.5987	.6026	.6064	.6103	.6141
.3	.6179	.6217	.6255	.6293	.6331	.6368	.6406	.6443	.6480	.6517
.4	.6554	.6591	.6628	.6664	.6700	.6736	.6772	.6808	.6844	.6879
.5	.6915	.6950	.6985	.7019	.7054	.7088	.7123	.7157	.7190	.7224
.6	.7257	.7291	.7324	.7357	.7389	.7422	.7454	.7486	.7517	.7549
.7	.7580	.7611	.7642	.7673	.7704	.7734	.7764	.7794	.7823	.7852
.8	.7881	.7910	.7939	.7967	.7995	.8023	.8051	.8078	.8106	.8133
.9	.8159	.8186	.8212	.8238	.8264	.8289	.8315	.8340	.8365	.8389
1.0	.8413	.8438	.8461	.8485	.8508	.8531	.8554	.8577	.8599	.8621
1.1	.8643	.8665	.8686	.8708	.8729	.8749	.8770	.8790	.8810	.8830
1.2	.8849	.8869	.8888	.8907	.8925	.8944	.8962	.8980	.8997	.9015
1.3	.9032	.9049	.9066	.9082	.9099	.9115	.9131	.9147	.9162	.9177
1.4	.9192	.9207	.9222	.9236	.9251	.9265	.9279	.9292	.9306	.9319
1.5	.9332	.9345	.9357	.9370	.9382	.9394	.9406	.9418	.9429	.9441
1.6	.9452	.9463	.9474	.9484	.9495	.9505	.9515	.9525	.9535	.9545
1.7	.9554	.9564	.9573	.9582	.9591	.9599	.9608	.9616	.9625	.9633
1.8	.9641	.9649	.9656	.9664	.9671	.9678	.9686	.9693	.9699	.9706
1.9	.9713	.9719	.9726	.9732	.9738	.9744	.9750	.9756	.9761	.9767
2.0	.9772	.9778	.9783	.9788	.9793	.9798	.9803	.9808	.9812	.9817
2.1	.9821	.9826	.9830	.9834	.9838	.9842	.9846	.9850	.9854	.9857
2.2	.9861	.9864	.9868	.9871	.9875	.9878	.9881	.9884	.9887	.9890
2.3	.9893	.9896	.9898	.9901	.9904	.9906	.9909	.9911	.9913	.9916
2.4	.9918	.9920	.9922	.9925	.9927	.9929	.9931	.9932	.9934	.9936

TABLE 3 Percentiles of the χ^2 Distribution—Values of χ^2_P Corresponding to P



df	$\chi^2_{.005}$	$\chi^2_{.01}$	$\chi^2_{.025}$	$\chi^2_{.05}$	$\chi^2_{.10}$	$\chi^2_{.90}$	$\chi^2_{.95}$	$\chi^2_{.975}$	$\chi^2_{.99}$	$\chi^2_{.995}$
1	.000039	.00016	.00098	.0039	.0158	2.71	3.84	5.02	6.63	7.88
2	.0100	.0201	.0506	.1026	.2107	4.61	5.99	7.38	9.21	10.60
3	.0717	.115	.216	.352	.584	6.25	7.81	9.35	11.34	12.84
4	.207	.297	.484	.711	1.064	7.78	9.49	11.14	13.28	14.86
5	.412	.554	.831	1.15	1.61	9.24	11.07	12.83	15.09	16.75
6	.676	.872	1.24	1.64	2.20	10.64	12.59	14.45	16.81	18.55
7	.989	1.24	1.69	2.17	2.83	12.02	14.07	16.01	18.48	20.28
8	1.34	1.65	2.18	2.73	3.49	13.36	15.51	17.53	20.09	21.96
9	1.73	2.09	2.70	3.33	4.17	14.68	16.92	19.02	21.67	23.59
10	2.16	2.56	3.25	3.94	4.87	15.99	18.31	20.48	23.21	25.19
11	2.60	3.05	3.82	4.57	5.58	17.28	19.68	21.92	24.73	26.76
12	3.07	3.57	4.40	5.23	6.30	18.55	21.03	23.34	26.22	28.30
13	3.57	4.11	5.01	5.89	7.04	19.81	22.36	24.74	27.69	29.82
14	4.07	4.66	5.63	6.57	7.79	21.06	23.68	26.12	29.14	31.32
15	4.60	5.23	6.26	7.26	8.55	22.31	25.00	27.49	30.58	32.80

TABLE 4 Percentiles of the t Distribution



df	$t_{.60}$	$t_{.70}$	$t_{.80}$	$t_{.90}$	$t_{.95}$	$t_{.975}$	$t_{.99}$	$t_{.995}$
1	.325	.727	1.376	3.078	6.314	12.706	31.821	63.657
2	.289	.617	1.061	1.886	2.920	4.303	6.965	9.925
3	.277	.584	.978	1.638	2.353	3.182	4.541	5.841
4	.271	.569	.941	1.533	2.132	2.776	3.747	4.604
5	.267	.559	.920	1.476	2.015	2.571	3.365	4.032
6	.265	.553	.906	1.440	1.943	2.447	3.143	3.707
7	.263	.549	.896	1.415	1.895	2.365	2.998	3.499
8	.262	.546	.889	1.397	1.860	2.306	2.896	3.355
9	.261	.543	.883	1.383	1.833	2.262	2.821	3.250
10	.260	.542	.879	1.372	1.812	2.228	2.764	3.169
11	.260	.540	.876	1.363	1.796	2.201	2.718	3.106
12	.259	.539	.873	1.356	1.782	2.179	2.681	3.055
13	.259	.538	.870	1.350	1.771	2.160	2.650	3.012
14	.258	.537	.868	1.345	1.761	2.145	2.624	2.977
15	.258	.536	.866	1.341	1.753	2.131	2.602	2.947