Post Exam 2 Practice Questions, 18.05, Spring 2022

Note: This is a set of practice problems for the material that came after exam 2. In preparing for the final you should use the previous review materials, the psets, the notes and slides.

1 Confidence intervals

To practice for the exam use the t and z-tables supplied at the end of this file. Be sure to learn to use these tables. Note the t and z-tables give left tail probabilities and the χ^2-table gives right tail critical values.

Problem 1. Basketball
Suppose that against a certain opponent the number of points the MIT basketball team scores is normally distributed with unknown mean θ and unknown variance, σ^2.

Suppose that over the course of the last 10 games between the two teams MIT scored the following points:

$$59, 62, 59, 74, 70, 61, 62, 66, 62, 75$$

(a) Compute a 95% t–confidence interval for θ. Does 95% confidence mean that the probability θ is in the interval you just found is 95%?

(b) Now suppose that you learn that $\sigma^2 = 25$. Compute a 95% z–confidence interval for θ. How does this compare to the interval in (a)?

(c) Let X be the number of points scored in a game. Suppose that your friend is a confirmed Bayesian with *a priori* belief $\theta \sim N(60, 16)$ and that $X \sim N(\theta, 25)$. He computes a 95% probability interval for θ, given the data in part (a). How does this interval compare to the intervals in (a) and (b)?

(d) Which of the three intervals constructed above do you prefer? Why?

Problem 2. Confidence interval 2
The volume in a set of wine bottles is known to follow a $\mathrm{N}(\mu, 25)$ distribution. You take a sample of the bottles and measure their volumes. How many bottles do you have to sample to have a 95% confidence interval for μ with width 1?

Problem 3. Suppose data $x_1, \dots, x_n$ are i.i.d. and drawn from $\mathrm{N}(\mu, \sigma^2)$, where μ and σ are unknown.

Suppose a data set is taken and we have $n = 49$, sample mean $\bar{x} = 92$ and sample standard deviation $s = 0.75$.

Find a 90% confidence interval for μ.

Problem 4. Polling confidence intervals
You do a poll to see what fraction p of the population supports candidate A over candidate B.

(a) How many people do you need to poll to know p to within 1% with 95% confidence.

(b) Let p be the fraction of the population who prefer candidate A. If you poll 400 people, how many have to prefer candidate A so that the 90% confidence interval is entirely above $p = 0.5$.

Problem 5. Confidence intervals 3
Suppose you made 40 confidence intervals with confidence level 95%. About how many of them would you expect to be “wrong’? That is, how many would not actually contain the parameter being estimated? Should you be surprised if 10 of them are wrong?

2 Chi-square confidence interval

Problem 6. Hotel
A statistician chooses 27 randomly selected dates, and when examining the occupancy records of a particular motel for those dates, finds a standard deviation of 5.86 rooms rented. If the number of rooms rented is normally distributed, find the 95% confidence interval for the population standard deviation of the number of rooms rented.

3 Bootstrapping

Problem 7. Parametric bootstrap
Suppose we have a sample of size 100 drawn from a geom(p) distribution with unknown p. The MLE estimate for p is given by by $\hat{p} = 1/\bar{x}$. Assume for our data $\bar{x} = 3.30$, so $\hat{p} = 1/\bar{x} = 0.30303$.

(a) Outline the steps needed to generate a parametric basic bootstrap 90% confidence interval.

(b) Suppose the following sorted list consists of 200 bootstrap means computed from a sample of size 100 drawn from a geometric(0.30303) distribution. Use the list to construct a 90% basic CI for p.

```
2.68 2.77 2.79 2.81 2.82 2.84 2.84 2.85 2.88 2.89
2.91 2.91 2.91 2.92 2.94 2.94 2.95 2.97 2.97 2.99
3.00 3.00 3.01 3.01 3.01 3.03 3.04 3.04 3.04 3.04
3.04 3.05 3.06 3.06 3.07 3.07 3.07 3.08 3.08 3.08
3.08 3.09 3.09 3.10 3.11 3.11 3.12 3.13 3.13 3.13
3.13 3.15 3.15 3.15 3.16 3.16 3.16 3.16 3.17 3.17
3.17 3.18 3.20 3.20 3.20 3.21 3.21 3.22 3.23 3.23
3.23 3.23 3.23 3.24 3.24 3.24 3.24 3.25 3.25 3.25
3.25 3.25 3.25 3.26 3.26 3.26 3.26 3.27 3.27 3.27
3.28 3.29 3.29 3.30 3.30 3.30 3.30 3.30 3.30 3.31
3.31 3.32 3.32 3.34 3.34 3.34 3.34 3.35 3.35 3.35
3.35 3.35 3.36 3.36 3.37 3.37 3.37 3.37 3.37 3.37
3.38 3.38 3.39 3.39 3.40 3.40 3.40 3.40 3.41 3.42
3.42 3.42 3.43 3.43 3.43 3.43 3.44 3.44 3.44 3.44
3.44 3.45 3.45 3.45 3.45 3.45 3.45 3.45 3.46 3.46
3.46 3.46 3.47 3.47 3.49 3.49 3.49 3.49 3.49 3.50
```

```
3.50 3.50 3.52 3.52 3.52 3.52 3.53 3.54 3.54 3.54
3.55 3.56 3.57 3.58 3.59 3.59 3.60 3.61 3.61 3.61
3.62 3.63 3.65 3.65 3.67 3.67 3.68 3.70 3.72 3.72
3.73 3.73 3.74 3.76 3.78 3.79 3.80 3.86 3.89 3.91
```

Problem 8. Empirical bootstrap
Suppose we had 100 data points $x_1, \ldots x_{100}$ with sample median $\hat{q_{0.5}} = 3.3$.

(a) Outline the steps needed to generate an empirical percentile bootstrap 90% confidence interval for the median $q_{0.5}$.

(b) Suppose now that the sorted list in the previous problem consists of 200 empirical bootstrap medians computed from resamples of size 100 drawn from the original data. Use the list to construct a 90% precentile CI for $q_{0.5}$.

4 Linear regression/Least squares

Problem 9. In this problem we will use maximum likelihood estimates to develop Gauss' method of least squares for fitting lines to data.

Suppose you have bivariate data, that is, a sequence of pairs $(x_1, y_1), \ldots, (x_n, y_n)$. A common model is that there is a linear relationship between x and y, so in principle the data should lie exactly along a line. However, since data has random noise this will not be the case. What we can do is look for the line that best fits the data. To do this we will use a model called simple linear regression.

For bivariate data the simple linear regression model assumes that the x_i are not random and that, for some values of the parameters a and b, we have

$$y_i = ax_i + b + \text{ random noise}$$

To be more precise, we will assume that the value y_i is drawn from a random variable of the form

$$Y_i \sim ax_i + b + \varepsilon_i$$

where ε_i is a normal random variable with mean 0 and variance σ^2. We assume all of the random variables ε_i are independent and that σ is the same for all i.

Notes. 1. The goal in simple linear regression is to find the line, i.e. the values of a and b that best fit the data.

2. We think of ε_i as the random measurement error.

3. Remember that (x_i, y_i) are not variables. They are data values.

(a) The distribution of Y_i depends on a, b, σ and x_i. So we write its pdf as

$$f(y_i \mid a, b, x_i, \sigma).$$

Give the formula for the likelihood function corresponding to one random value y_i. (Hint: $y_i - ax_i - b \sim \mathrm{N}(0, \sigma^2)$.)

(b) For general data (x_1, y_1), ..., (x_n, y_n) give the likelihood and log likelihood functions (again as functions of a, b, and σ).

(c) Find the maximum likelihood estimate for a and b by taking partial derivatives of the log likelihood function and setting them equal to 0.

(d) Suppose we have data $(1, 8)$, $(3, 2)$, $(5, 1)$. Use your answer in part (b) to find the value of a and b which gives the MLE for the best fitting line to the data.

(e) Use R to plot the data and the regression line you found in problem (1c). The commands `plot(x, y, pch=19)` and `abline()` will come in handy.

Print the plot and turn it in.

Problem 10. What is the relationship between correlation and least squares fit line?

Problem 11. You have bivariate data (x_i, y_i). You have reason to suspect the data is related by $y_i = a/x_i + U_i$ where U_i is a random variable with mean 0 and variance σ^2 (the same for all i).

Find the least squares estimate of a.

Problem 12. Least Squares and MLE

In this problem we will see that the least squares fit of a line is just the MLE assuming the error terms are normally distributed.

For bivariate data $(x_1, y_1), \ldots, (x_n, y_n)$, the simple linear regression model says that y_i is a random value generated by a random variable

$$Y_i = ax_i + b + \varepsilon_i$$

where a, b, x_i are fixed (not random) values, and ε_i is a random variable with mean 0 and variance σ^2.

(a) Suppose that each $\varepsilon_i \sim \mathrm{N}(0, \sigma^2)$. Show that $Y_i \sim \mathrm{N}(ax_i + b, \sigma^2)$.

(b) Give the formula for the pdf $f_{Y_i}(y_i)$ of Y_i.

(c) Write down the likelihood of the data as a function of a, b, and σ.

Standard normal table of left tail probabilities.

z	$\Phi(z)$	z	$\Phi(z)$	z	$\Phi(z)$	z	$\Phi(z)$
-4.00	0.0000	-2.00	0.0228	0.00	0.5000	2.00	0.9772
-3.95	0.0000	-1.95	0.0256	0.05	0.5199	2.05	0.9798
-3.90	0.0000	-1.90	0.0287	0.10	0.5398	2.10	0.9821
-3.85	0.0001	-1.85	0.0322	0.15	0.5596	2.15	0.9842
-3.80	0.0001	-1.80	0.0359	0.20	0.5793	2.20	0.9861
-3.75	0.0001	-1.75	0.0401	0.25	0.5987	2.25	0.9878
-3.70	0.0001	-1.70	0.0446	0.30	0.6179	2.30	0.9893
-3.65	0.0001	-1.65	0.0495	0.35	0.6368	2.35	0.9906
-3.60	0.0002	-1.60	0.0548	0.40	0.6554	2.40	0.9918
-3.55	0.0002	-1.55	0.0606	0.45	0.6736	2.45	0.9929
-3.50	0.0002	-1.50	0.0668	0.50	0.6915	2.50	0.9938
-3.45	0.0003	-1.45	0.0735	0.55	0.7088	2.55	0.9946
-3.40	0.0003	-1.40	0.0808	0.60	0.7257	2.60	0.9953
-3.35	0.0004	-1.35	0.0885	0.65	0.7422	2.65	0.9960
-3.30	0.0005	-1.30	0.0968	0.70	0.7580	2.70	0.9965
-3.25	0.0006	-1.25	0.1056	0.75	0.7734	2.75	0.9970
-3.20	0.0007	-1.20	0.1151	0.80	0.7881	2.80	0.9974
-3.15	0.0008	-1.15	0.1251	0.85	0.8023	2.85	0.9978
-3.10	0.0010	-1.10	0.1357	0.90	0.8159	2.90	0.9981
-3.05	0.0011	-1.05	0.1469	0.95	0.8289	2.95	0.9984
-3.00	0.0013	-1.00	0.1587	1.00	0.8413	3.00	0.9987
-2.95	0.0016	-0.95	0.1711	1.05	0.8531	3.05	0.9989
-2.90	0.0019	-0.90	0.1841	1.10	0.8643	3.10	0.9990
-2.85	0.0022	-0.85	0.1977	1.15	0.8749	3.15	0.9992
-2.80	0.0026	-0.80	0.2119	1.20	0.8849	3.20	0.9993
-2.75	0.0030	-0.75	0.2266	1.25	0.8944	3.25	0.9994
-2.70	0.0035	-0.70	0.2420	1.30	0.9032	3.30	0.9995
-2.65	0.0040	-0.65	0.2578	1.35	0.9115	3.35	0.9996
-2.60	0.0047	-0.60	0.2743	1.40	0.9192	3.40	0.9997
-2.55	0.0054	-0.55	0.2912	1.45	0.9265	3.45	0.9997
-2.50	0.0062	-0.50	0.3085	1.50	0.9332	3.50	0.9998
-2.45	0.0071	-0.45	0.3264	1.55	0.9394	3.55	0.9998
-2.40	0.0082	-0.40	0.3446	1.60	0.9452	3.60	0.9998
-2.35	0.0094	-0.35	0.3632	1.65	0.9505	3.65	0.9999
-2.30	0.0107	-0.30	0.3821	1.70	0.9554	3.70	0.9999
-2.25	0.0122	-0.25	0.4013	1.75	0.9599	3.75	0.9999
-2.20	0.0139	-0.20	0.4207	1.80	0.9641	3.80	0.9999
-2.15	0.0158	-0.15	0.4404	1.85	0.9678	3.85	0.9999
-2.10	0.0179	-0.10	0.4602	1.90	0.9713	3.90	1.0000
-2.05	0.0202	-0.05	0.4801	1.95	0.9744	3.95	1.0000

$\Phi(z) = P(Z \le z)$ for N$(0,1)$.

(Use interpolation to estimate z values to a 3rd decimal place.)

Table of Student t critical values (right-tail)

The table shows $t_{df,p}$ = the $1-p$ quantile of $t(df)$.

We only give values for $p \leq 0.5$. Use symmetry to find the values for $p > 0.5$, e.g.

$$t_{5,\,0.975} = -t_{5,\,0.025}$$

In R notation $t_{df,p}$ = `qt(1-p, df)`.

df\p	0.005	0.010	0.015	0.020	0.025	0.030	0.040	0.050	0.100	0.200	0.300	0.400	0.500
1	63.66	31.82	21.20	15.89	12.71	10.58	7.92	6.31	3.08	1.38	0.73	0.32	0.00
2	9.92	6.96	5.64	4.85	4.30	3.90	3.32	2.92	1.89	1.06	0.62	0.29	0.00
3	5.84	4.54	3.90	3.48	3.18	2.95	2.61	2.35	1.64	0.98	0.58	0.28	0.00
4	4.60	3.75	3.30	3.00	2.78	2.60	2.33	2.13	1.53	0.94	0.57	0.27	0.00
5	4.03	3.36	3.00	2.76	2.57	2.42	2.19	2.02	1.48	0.92	0.56	0.27	0.00
6	3.71	3.14	2.83	2.61	2.45	2.31	2.10	1.94	1.44	0.91	0.55	0.26	0.00
7	3.50	3.00	2.71	2.52	2.36	2.24	2.05	1.89	1.41	0.90	0.55	0.26	0.00
8	3.36	2.90	2.63	2.45	2.31	2.19	2.00	1.86	1.40	0.89	0.55	0.26	0.00
9	3.25	2.82	2.57	2.40	2.26	2.15	1.97	1.83	1.38	0.88	0.54	0.26	0.00
10	3.17	2.76	2.53	2.36	2.23	2.12	1.95	1.81	1.37	0.88	0.54	0.26	0.00
16	2.92	2.58	2.38	2.24	2.12	2.02	1.87	1.75	1.34	0.86	0.54	0.26	0.00
17	2.90	2.57	2.37	2.22	2.11	2.02	1.86	1.74	1.33	0.86	0.53	0.26	0.00
18	2.88	2.55	2.36	2.21	2.10	2.01	1.86	1.73	1.33	0.86	0.53	0.26	0.00
19	2.86	2.54	2.35	2.20	2.09	2.00	1.85	1.73	1.33	0.86	0.53	0.26	0.00
20	2.85	2.53	2.34	2.20	2.09	1.99	1.84	1.72	1.33	0.86	0.53	0.26	0.00
21	2.83	2.52	2.33	2.19	2.08	1.99	1.84	1.72	1.32	0.86	0.53	0.26	0.00
22	2.82	2.51	2.32	2.18	2.07	1.98	1.84	1.72	1.32	0.86	0.53	0.26	0.00
23	2.81	2.50	2.31	2.18	2.07	1.98	1.83	1.71	1.32	0.86	0.53	0.26	0.00
24	2.80	2.49	2.31	2.17	2.06	1.97	1.83	1.71	1.32	0.86	0.53	0.26	0.00
25	2.79	2.49	2.30	2.17	2.06	1.97	1.82	1.71	1.32	0.86	0.53	0.26	0.00
30	2.75	2.46	2.28	2.15	2.04	1.95	1.81	1.70	1.31	0.85	0.53	0.26	0.00
31	2.74	2.45	2.27	2.14	2.04	1.95	1.81	1.70	1.31	0.85	0.53	0.26	0.00
32	2.74	2.45	2.27	2.14	2.04	1.95	1.81	1.69	1.31	0.85	0.53	0.26	0.00
33	2.73	2.44	2.27	2.14	2.03	1.95	1.81	1.69	1.31	0.85	0.53	0.26	0.00
34	2.73	2.44	2.27	2.14	2.03	1.95	1.80	1.69	1.31	0.85	0.53	0.26	0.00
35	2.72	2.44	2.26	2.13	2.03	1.94	1.80	1.69	1.31	0.85	0.53	0.26	0.00
40	2.70	2.42	2.25	2.12	2.02	1.94	1.80	1.68	1.30	0.85	0.53	0.26	0.00
41	2.70	2.42	2.25	2.12	2.02	1.93	1.80	1.68	1.30	0.85	0.53	0.25	0.00
42	2.70	2.42	2.25	2.12	2.02	1.93	1.79	1.68	1.30	0.85	0.53	0.25	0.00
43	2.70	2.42	2.24	2.12	2.02	1.93	1.79	1.68	1.30	0.85	0.53	0.25	0.00
44	2.69	2.41	2.24	2.12	2.02	1.93	1.79	1.68	1.30	0.85	0.53	0.25	0.00
45	2.69	2.41	2.24	2.12	2.01	1.93	1.79	1.68	1.30	0.85	0.53	0.25	0.00
46	2.69	2.41	2.24	2.11	2.01	1.93	1.79	1.68	1.30	0.85	0.53	0.25	0.00
47	2.68	2.41	2.24	2.11	2.01	1.93	1.79	1.68	1.30	0.85	0.53	0.25	0.00
48	2.68	2.41	2.24	2.11	2.01	1.93	1.79	1.68	1.30	0.85	0.53	0.25	0.00
49	2.68	2.40	2.24	2.11	2.01	1.93	1.79	1.68	1.30	0.85	0.53	0.25	0.00

Table of χ^2 critical values (right-tail)

The table shows $c_{df,p}$ = the $1-p$ quantile of $\chi^2(df)$.

In R notation $c_{df,p}$ = `qchisq(1-p, df)`.

df\p	0.010	0.025	0.050	0.100	0.200	0.300	0.500	0.700	0.800	0.900	0.950	0.975	0.990
1	6.63	5.02	3.84	2.71	1.64	1.07	0.45	0.15	0.06	0.02	0.00	0.00	0.00
2	9.21	7.38	5.99	4.61	3.22	2.41	1.39	0.71	0.45	0.21	0.10	0.05	0.02
3	11.34	9.35	7.81	6.25	4.64	3.66	2.37	1.42	1.01	0.58	0.35	0.22	0.11
4	13.28	11.14	9.49	7.78	5.99	4.88	3.36	2.19	1.65	1.06	0.71	0.48	0.30
5	15.09	12.83	11.07	9.24	7.29	6.06	4.35	3.00	2.34	1.61	1.15	0.83	0.55
6	16.81	14.45	12.59	10.64	8.56	7.23	5.35	3.83	3.07	2.20	1.64	1.24	0.87
7	18.48	16.01	14.07	12.02	9.80	8.38	6.35	4.67	3.82	2.83	2.17	1.69	1.24
8	20.09	17.53	15.51	13.36	11.03	9.52	7.34	5.53	4.59	3.49	2.73	2.18	1.65
9	21.67	19.02	16.92	14.68	12.24	10.66	8.34	6.39	5.38	4.17	3.33	2.70	2.09
10	23.21	20.48	18.31	15.99	13.44	11.78	9.34	7.27	6.18	4.87	3.94	3.25	2.56
16	32.00	28.85	26.30	23.54	20.47	18.42	15.34	12.62	11.15	9.31	7.96	6.91	5.81
17	33.41	30.19	27.59	24.77	21.61	19.51	16.34	13.53	12.00	10.09	8.67	7.56	6.41
18	34.81	31.53	28.87	25.99	22.76	20.60	17.34	14.44	12.86	10.86	9.39	8.23	7.01
19	36.19	32.85	30.14	27.20	23.90	21.69	18.34	15.35	13.72	11.65	10.12	8.91	7.63
20	37.57	34.17	31.41	28.41	25.04	22.77	19.34	16.27	14.58	12.44	10.85	9.59	8.26
21	38.93	35.48	32.67	29.62	26.17	23.86	20.34	17.18	15.44	13.24	11.59	10.28	8.90
22	40.29	36.78	33.92	30.81	27.30	24.94	21.34	18.10	16.31	14.04	12.34	10.98	9.54
23	41.64	38.08	35.17	32.01	28.43	26.02	22.34	19.02	17.19	14.85	13.09	11.69	10.20
24	42.98	39.36	36.42	33.20	29.55	27.10	23.34	19.94	18.06	15.66	13.85	12.40	10.86
25	44.31	40.65	37.65	34.38	30.68	28.17	24.34	20.87	18.94	16.47	14.61	13.12	11.52
30	50.89	46.98	43.77	40.26	36.25	33.53	29.34	25.51	23.36	20.60	18.49	16.79	14.95
31	52.19	48.23	44.99	41.42	37.36	34.60	30.34	26.44	24.26	21.43	19.28	17.54	15.66
32	53.49	49.48	46.19	42.58	38.47	35.66	31.34	27.37	25.15	22.27	20.07	18.29	16.36
33	54.78	50.73	47.40	43.75	39.57	36.73	32.34	28.31	26.04	23.11	20.87	19.05	17.07
34	56.06	51.97	48.60	44.90	40.68	37.80	33.34	29.24	26.94	23.95	21.66	19.81	17.79
35	57.34	53.20	49.80	46.06	41.78	38.86	34.34	30.18	27.84	24.80	22.47	20.57	18.51
40	63.69	59.34	55.76	51.81	47.27	44.16	39.34	34.87	32.34	29.05	26.51	24.43	22.16
41	64.95	60.56	56.94	52.95	48.36	45.22	40.34	35.81	33.25	29.91	27.33	25.21	22.91
42	66.21	61.78	58.12	54.09	49.46	46.28	41.34	36.75	34.16	30.77	28.14	26.00	23.65
43	67.46	62.99	59.30	55.23	50.55	47.34	42.34	37.70	35.07	31.63	28.96	26.79	24.40
44	68.71	64.20	60.48	56.37	51.64	48.40	43.34	38.64	35.97	32.49	29.79	27.57	25.15
45	69.96	65.41	61.66	57.51	52.73	49.45	44.34	39.58	36.88	33.35	30.61	28.37	25.90
46	71.20	66.62	62.83	58.64	53.82	50.51	45.34	40.53	37.80	34.22	31.44	29.16	26.66
47	72.44	67.82	64.00	59.77	54.91	51.56	46.34	41.47	38.71	35.08	32.27	29.96	27.42
48	73.68	69.02	65.17	60.91	55.99	52.62	47.34	42.42	39.62	35.95	33.10	30.75	28.18
49	74.92	70.22	66.34	62.04	57.08	53.67	48.33	43.37	40.53	36.82	33.93	31.55	28.94