Mathematical Statistics: Homework Problems and Experimental Project (Spring 2024)[1]

General guideline. Use any help you want, including people, computer algebra systems, Internet, and solution manuals, but make sure you are ready for quizzes and exams, where you are on your own.

The experimental project. [The project is due on Friday, April 26 (upload to Blackboard as a **single PDF file**).] *Your task is to carry our statistical analysis of some real-life data.* Here is a general outline:

(1) Provide suitable background about the problem.
(2) Describe the assumptions you make about the probability distributions involved in the problem.
(3) State the appropriate null and alternative hypothesis you are planning to test.
(4) Design an experiment to test the hypothesis. Keeping in mind the ideas from the book [e.g. Chapter 12] write a short paragraph describing the experiment and explaining why you think your design is reasonable.
(5) Conduct the experiment and organize the data you collect.
(6) Use a suitable statistical test to test your hypothesis.
(7) Compute the p-value and explain whether the null hypothesis should be rejected or not.
(8) Write an overall summary.

The final product should be a [reasonable well written and presented] report you wrote yourself, with a reasonably complete understanding of all the details, *so that I should be able to get the main idea within 1 min of looking through it, and you should be able to understand it 5 years from now.* Most importantly, you yourself should like what you did.

Both external help and internal collaboration are encouraged.

`This project is your chance to show your creativity, in particular, by finding an interesting question to investigate.` First and foremost, you should find the question interesting for yourself. If you are running out of time or out of ideas [or just feel lazy...] here are two questions that I find interesting:

(1) Does a shuffle function on a digital music player produce a random permutation? Make a list of 10 (or more...) songs and play them at random *without replacement*. This will produce a permutation of the list. Is every permutation equally likely to be produced? [Notice that you might not have enough time to sample all 10! permutations, but you can try...it is a challenge in itself...]
(2) Pick your favorite `irrational` number, such as $\sqrt{2}$, π, e, ϕ, or γ. Do the digits in the decimal expansion of this number appear in random order? In other words, as you get more and more digits in the expansion, is every of the ten digits $0, 1, \ldots, 9$ equally likely to appear? What about a particular sequence of length 2, or 3, or 4, etc.? For example, is 1324 more likely than 8888?

Homework 1.

Problem 1. Determine the mean and the standard deviation of a normal random variable X in each of the two cases:

$$\text{Case 1}: \ P(X>0)=\frac{1-0.5763}{2} \quad \text{and} \quad P(X<2)=\frac{1+0.8664}{2};$$
$$\text{Case 2}: \ P(X<0)=\frac{1-0.5763}{2} \quad \text{and} \quad P(X<2)=\frac{1+0.8664}{2}.$$

Note. The numbers 0.5763 and 0.8664 come directly from a Z-table, and should lead you to the right Z-value. The Z-value corresponding to 0.5763 is ± 0.8; to figure out whether you take positive or negative value, draw a picture.

Problem 2. Let Z be a standard Gaussian random variable. Determine the values of the real number r for which $\mathbb{E}|Z|^r$ exits and compute the expectation for those r. Express your

[1]Sergey Lototsky, USC

answer in terms of the Gamma function

$$\Gamma(x) = \int_0^{+\infty} t^{x-1} e^{-t} dt,$$

and simplify the answer when possible (for example, when r is a positive even integer). In particular, confirm that $\mathbb{E}Z^4 = 3$.

Problem 3. Let X and Y be independent standard Gaussian random variables and define two more random variables: $U = X+2Y$, $V = 3X-2Y$. Compute the conditional expectation $W = \mathbb{E}(V|U)$ [it should be $-U/5$]and confirm that $V - W$ and W are independent [easy to see because they are uncorrelated and jointly Gaussian: write them both in terms of X and Y].

For best results, do this problem two ways: (a) by using the Normal Correlation Theorem and (b) by using vectors in the plane, thinking of X and Y as the vectors $\hat{\imath}$ and $\hat{\jmath}$ so that $U = \langle 1, 2\rangle$, $V = \langle 3, -2\rangle$, and W is the orthogonal projection of V on U.

Problem 4. Recall that the random variable Y has χ^2_n distribution if $Y = X_1^2 + \cdots + X_n^2$, where X_k, $k \geq 1$, are independent identically distributed (iid) standard Gaussian random variables. Compute the mode, mean, and variance of Y. Sketch the graph of the pdf of Y for $n = 1, 2, 3, 4$ and $n \geq 5$. [The back cover of the book provides all the missing information].

Problem 5. A certain town has 250,000 families, of which 25,000 do not have a TV at home. As part of an opinion survey, a simple random sample of 900 families is chosen. What is the chance that between 9% and 11% of the sample families will not have a TV at home? [This is a straightforward problem on the CLT for binomial distribution; the answer is $P(|Z| < 1) \approx 0.68$]

Homework 2.

Problem 1. Given the set of numbers

0.24956	0.35335	0.13951	0.57409	0.37571
0.74346	0.33675	0.21273	0.28307	0.75326
0.5015	0.59701	0.34005	0.12086	0.4583
0.35671	0.77455	0.16124	0.46059	0.16075
0.10534	0.82210	0.52970	0.47606	0.39321

compute the following: (a) the sample range, (b) the sample mean $\bar{X}$, (c) the sample median, (d) the sample standard deviations $\bar{\sigma}$ and s, (e) the sample skewness and sample kurtosis, (f) D_1, the average of $|X - \bar{X}|$, and D_4, the fourth root of the average of $(X - \bar{X})^4$, (g) confirm that $D_1 < \bar{\sigma} = D_2 < D_4$ and explain why the inequalities must hold.

Problem 2. Let $X_1, \ldots, X_n$ be iid with mean μ and variance σ^2, and let

$$\bar{X}_n = \frac{1}{n} \sum_{k=1}^{n} X_k.$$

(a) Confirm that

$$\bar{\sigma}_n^2 = \frac{1}{n} \sum_{k=1}^{n} (X_k - \bar{X}_n)^2$$

is a biased estimator of σ^2, whereas

$$\tilde{\sigma}_n^2 = \frac{1}{n} \sum_{k=1}^{n} (X_k - \mu)^2 \quad \text{and} \quad s_n^2 = \frac{1}{n-1} \sum_{k=1}^{n} (X_k - \bar{X}_n)^2$$

are unbiased estimators of σ^2.

(b) In the case of the normal distribution, compute the variance and the MSE of each of the three estimators. Keep in mind that now $n\tilde{\sigma}_n^2/\sigma^2$ has χ^2_n distribution, and $(n-1)s_n^2/\sigma^2$ has

χ^2_{n-1} distribution. The mean of χ^2_n is n, the variance is $2n$. This can speed up computations so that you do not need to expand either $\tilde{\sigma}^2_n$ or s^2_n. For example,

$$\mathbb{E}\big(s^2_n - \sigma^2\big)^2 = \frac{\sigma^4}{(n-1)^2}\mathbb{E}\big((n-1)s^2_n/\sigma^2 - (n-1)\big)^2 = \frac{\sigma^4}{(n-1)^2}\mathrm{Var}\big(\chi^2_{n-1}\big) = \frac{2\sigma^4}{n-1}.$$

Problem 3. Let $X_1, \ldots, X_n$ be a random sample from a population with pdf $f = f(x;\theta)$. In each of the three cases below, confirm that the given function is indeed a pdf, compute the MSE of the given estimator $\widehat{\theta}_n$ of θ, and check whether the MSE goes to zero as $n \to \infty$:

Case 1 : $f(x;\theta) = a\theta^{-a}x^{a-1},\ a > 0,\ 0 < x < \theta;\ \ \widehat{\theta}_n = X_{(n)} = \max(X_1, \ldots, X_n);$

Case 2 : $f(x;\theta) = a\theta^{a}x^{-a-1},\ a > 0,\ 0 < \theta < x;\ \ \widehat{\theta}_n = X_{(1)} = \min(X_1, \ldots, X_n);$

Case 3 : $f(x;\theta) = \theta^{-1}e^{-x/\theta},\ x > 0,\ \theta > 0;\ \widehat{\theta}_n = nX_{(1)}.$

Homework 3.

Problem 1. The lifetime of a toaster from the company Toaster's Choice has a normal distribution with standard deviation 1.5 years. A random sample of 400 toasters was drawn yielding the sample lifetime average of 6 years.

a) Compute a 90% confidence interval for the mean lifetime of the toasters.
b) What sample size is needed to find the mean lifetime of the toasters to within plus or minus 0.05 years at the same 90% confidence level?
c) How will the answers in parts a) and b) change if, instead of knowing the standard deviation to be 1.5 years, it was estimated to be 1.5 years, based on the same sample of 400 devices.
d) Do parts a) and b) under the assumption that the lifetime has normal distribution, but with unknown standard deviation, and that a sample of 10 devices produced sample lifetime average 6 years and sample standard deviation s_{10} 1.5 years.
e) Compare the intervals from parts a) and d). Which one is longer? Does it make sense? Why?
f) Compare the sample sizes in parts b) and d). Which one is larger? Does it make sense? Why?

Problem 2. The ages of a random sample of five professors at a certain university are 39, 54, 61, 72, and 59. Assuming that the age of the professors in this university is normally distributed, construct the 95% confidence intervals for the mean and the standard deviation of the age.

Problem 3. In 1970, 59% of college freshmen thought that capital punishment should be abolished; in 2005, the percentage was 35%. The percentages are based on two independent simple random samples, each of size 1,000. Compute a 95% confidence interval for the difference in the percentages.

Problem 4. A study reports that freshmen at public universities work 10.2 hours a week for pay, on average, and the s_n is 8.5 hours; at private universities, the average is 8.1 hours and the s_n is 6.9 hours. Assume these data are based on two independent simple random samples, each of size 1,000. Construct a 95% confidence interval for the difference of the hours worked.

Problem 5. Let $X_1, \ldots, X_m$ be a random sample from a normal population with unknown mean μ_1 and unknown variance σ^2, and let $Y_1, \ldots, Y_n$ be an independent random sample from a normal population with unknown mean μ_2 and variance $k\sigma^2$, where $k > 0$ is known. Construct a $100(1-\alpha)\%$ confidence interval for $\mu_1 - \mu_2$. [You might be able to find a more detailed outline of the solution in the supplementary exercises at the end of Chapter 8.]

Homework 4.

The general framework for constructing problems on properties of estimators is as follows. Let $X_1, \dots, X_n$ be a random sample from a population, and the distribution of the population is characterized either by a pdf or a probability mass function $f(x;\theta)$.

(1) Construct a method-of-moments estimator of θ;
(2) Construct the MLE of θ;
(3) Determine a sufficient statistic for θ and construct an MVUE of θ;
(4) Given an estimator, compute its MSE, investigate its consistency, and compute its efficiency relative to some other estimator.
(5) If the distribution of the MLE, when appropriately normalized, is approximately standard normal [which happens often, but not always...], then you can also construct an approximate confidence interval for MLE.

You are encouraged to follow this guideline and make and solve as many problems as possible. Below is the bare minimum.

Problem 1. Let $f(x;\theta) = \theta^{-1}e^{-x/\theta},\ \theta > 0, x > 0$. Confirm that the sample mean is a consistent estimator of θ, and it coincides with the method-of-moments estimator, MLE, and MVUE. Compute its efficiency relative to $\widehat{\theta}_n = nX_{(1)} = n\min(X_1, \dots, X_n)$.

Problem 2. Assume that the population is Poisson with mean value θ. Confirm that the sample mean $\bar{X}_n$ is a consistent estimator of θ, and it coincides with the method-of-moments estimator, MLE, and MVUE.

Problem 3. Assume that $f(x;\theta) = e^{-(x-\theta)},\ x > \theta,\ \theta \in \mathbb{R}$. Confirm that $X_{(1)}$ is the MLE of θ and $X_{(1)} - (1/n)$ is an unbiased estimator of θ with variance $1/n^2$. Is $X_{(1)} - (1/n)$ an MVUE of θ?

Problem 4. Assume that the population is normal with mean μ and variance σ^2. Compute the method-of-moments estimator, MLE, and MVUE of the pair (μ, σ^2). Then think how the answers will change if one of the two numbers μ, σ^2 is known.

Problem 5. Assume that the population is Poisson with mean value $\theta > 0$. Confirm that $\sqrt{n}(\bar{X}_n - \theta)/\sqrt{\bar{X}_n}$ converges in distribution to Z and use the result to construct a 95% confidence interval for θ.

Homework 5.

Problem 1. Assume that the population is uniform on the interval $(0, \theta)$, $\theta > 0$.

(1) Construct the method-of-moments and the MLE of θ. Is MLE asymptotically normal?
(2) Confirm that $X_{(n)}$ is a sufficient statistic for θ and construct the MVUE of θ.
(3) Compute the efficiency of MVUE relative to the method-of-moments estimator.
(4) Confirm that the MSE of the estimator $\tilde{\theta}_n = (n+2)X_{(n)}/(n+1)$ is $\theta^2/(n+1)^2$, which is smaller than the MSE of the MVUE. How is this possible?

[Some answers: method of moments gives $2\bar{X}_n$, MLE is $X_{(n)}$, MVUE is $(n+1)X_{(n)}/n$ and its variance (which is the same as the MSE) is $\theta^2/(n(n+2))$; $\tilde{\theta}_n$ is biased.]

Problem 2. In 1970, 59% of college freshmen thought that capital punishment should be abolished; by 2005, the percentage had dropped to 35%. Is the difference real, or can it be explained by chance? Will your answer change if instead of difference, you look at the decrease? You may assume that the percentages are based on two independent simple random samples, each of size 1,000.

Problem 3. A study reports that freshmen at public universities work 10.2 hours a week for pay, on average, and the s_n is 8.5 hours; at private universities, the average is 8.1 hours and the s_n is 6.9 hours. Assume these data are based on two independent simple random

samples, each of size 1,000. Is the difference between the averages due to chance? If not, what else might explain it?

Problem 4. Suppose that the distribution of the test statistic to test the null hypothesis $a = 0$ against the alternative $a = 1/2$ is $f_0(x) = 2(1-x),\ 0 < x < 1$, under the null hypothesis and $f_1(x) = 2x,\ 0 < x < 1$, under the alternative. Suppose that the critical region is $[c, 1]$ and the observed value of the test statistic is y, where c and y are numbers between 0 and 1.

Compute the p-value of the experiment and the power of the test as functions of y and sketch the corresponding graphs. What do you expect from the power and p-value as $y \to 0$? As $y \to 1$? Do you expect the functions to be monotone? Are you getting the behavior you expect?

Problem 5. Assume that the distribution of the test statistic under the null hypothesis $\theta = \theta_0$ is symmetric around the origin [e.g. if you have a pdf, then it is an even function]. Confirm that the p-value in the case of the two-tail alternative ($\theta \neq \theta_0$) is twice the p-value of the corresponding upper-tail ($\theta > \theta_0$) or lower-tail ($\theta < \theta_0$) alternative.

Homework 6.

Problem 1. A market researcher for a consumer electronics company wants to determine if the residents of a particular city are spending more time watching TV than the average for this geographic area. The average for this geographic area is 13 hours per week. A random sample of 16 respondents of the city is selected, and each respondent is instructed to keep a detailed record of all television viewing in a particular week. For this sample the viewing time per week has a mean of 15.3 hours and a sample standard deviation $s_n = 3.8$ hours. Assume that the amount of time of television viewing per week is normally distributed. Can the researcher claim that the residents of this particular city are spending significantly (at 5% level) more time watching TV than the average for this geographic area? Explain your conclusion.

Problem 2. A study reports that freshmen at four-year public universities work 10.2 hours a week for pay, on average, and the s_n is 8.5 hours; at two-year community colleges, the average is 11.5 hours and the s_n is 8.5 hours. Assume these data are based on two independent simple random samples, each of size 16. Is the **difference** between the weekly work hours statistically significant (at 5% level)?

Problem 3. The standard deviation of the scores on an aptitude test is supposed to be high so that it is easier to distinguish between people with different abilities. Assume that the scores on a certain aptitude test are known to have standard deviation equal to 10. A new test is proposed and is tried on 20 people, producing the sample standard deviation of scores equal to 12. At what levels can you claim that the new test is significantly better?

Problem 4. Let $X_1, \dots, X_n$ be a random sample from the Gamma distribution with parameters $\alpha = 3$ and $\beta = \theta$. [Recall that the Gamma pdf is proportional to $x^{\alpha-1}e^{-x/\beta}$ and the sum of iid Gammas is again Gamma with the same beta-parameter]. Construct the likelihood ratio test of $H_0 : \theta = \theta_0$ against $H_1 : \theta > \theta_0$. [The answer should involve χ^2_{6n}, which is the distribution of $2n\bar{X}_n/\theta$]. Confirm that the test is uniformly most powerful against every alternative of the form $\theta = \theta_1$ with $\theta_1 > \theta_0$. [Here, you can either apply the Karlin-Rubin theorem or re-do the computations in the setting of the Neyman-Pearson lemma.]

Problem 5. Repeat the previous problem when the sample is from Poisson distribution with mean value $\lambda = \theta$.

Homework 7.

Problem 1. Recall a somewhat mysterious theorem saying that, under some regularity conditions, the random variable $-2\ln\lambda_n$ converges in distribution, as sample size n goes to

infinity, to a χ^2 random variable; here, λ_n is the test statistic in the likelihood ratio test. It makes sense to convince ourselves that the result is true in the most basic setting: testing for the normal mean.

Let $X_1, \dots, X_n$ be a random sample from a normal distribution with unknown variance, the null hypothesis is that the population mean is zero, and the alternative is that the population mean is not zero. Confirm that in this case

$$\lambda_n = \left(\frac{\sum_{k=1}^n X_k^2 - n\big(\bar{X}_n\big)^2}{\sum_{k=1}^n X_k^2} \right)^{n/2}$$

[you should be able to find most of the computations in the book] and, as $n \to \infty$, $-2\ln\lambda$ converges in distribution to χ_1^2 [here, with no loss of generality assume that $\sigma = 1$ and use that $\ln(1-x) \approx -x$ for x near 0, $\chi_1^2 = Z^2$, $\sqrt{n}\bar{X}_n = Z$, and, by the LLN, $\sum_{k=1}^n X_k^2 \approx n$].

Problem 2. Compute the least-squares estimate of the coefficient a in the *zero intercept* model $y_i = ax_i + \varepsilon_i$, $i = 1, \dots, n$.

Problem 3. Assume that $Y_i,\ i = 1, \dots, n$ are independent $\mathcal{N}(\beta_0 + x_i\beta_1, \sigma^2)$, with known x_i and unknown $\beta_0, \beta_1, \sigma^2$. Compute the maximum likelihood estimators for β_0, β_1, and σ^2. [For β_0 and β_1 you get the same estimators as with least squares.]

Problem 4. Introduce the column vectors $\vec{\theta} = (\beta_0, \beta_1)^\top$, $\vec{Y} = (Y_1, \dots, Y_n)^\top$, and $\vec{\varepsilon} = (\varepsilon_1, \dots, \varepsilon_n)^\top$. Define the matrix $G \in \mathbb{R}^{n\times 2}$ with rows $(1, x_1), \dots, (1, x_n)$. Confirm that `simple linear regression` model $Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i,\ i = 1, \dots, n$, can be written in the matrix-vector form

$$\vec{Y} = G\vec{\theta} + \vec{\varepsilon},$$

and

$$(\widehat{\beta}_0, \widehat{\beta}_1)^\top = (G^\top G)^{-1} G^\top \vec{Y}$$

provided $n \geq 2$. Conclude that if ε_i are iid $\mathcal{N}(0, \sigma^2)$ then the vector $(\widehat{\beta}_0, \widehat{\beta}_1)^\top$ is bivariate normal with mean $\vec{\theta}$ and covariance matrix $\sigma^2(G^\top G)^{-1}$.

Problem 5. Given the simple linear regression model $Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i,\ i = 1, \dots, n$, define the `residuals` $R_i = Y_i - (\widehat{\beta}_0 + \widehat{\beta}_1 x_i),\ i = 1, \dots, n$.

(a) Confirm that $\sum_{i=1}^n R_i = 0$.

(b) What if $\beta_0 = 0$ [zero-slope model]: is it still true that $\sum_{i=1}^n R_i = 0$?

Homework 8.

Problem 1. Let $Y_i,\ i = 1, \dots, n$ be independent $\mathcal{N}(\beta_0 + y_i\beta_1, \sigma^2)$ and let $W_1, \dots, W_m$ be independent $\mathcal{N}(\gamma_0 + w_i\gamma_1, \sigma^2)$, with known, and non-random, numbers $y_1, \dots, y_n, w_1, \dots, w_n$. Construct a test of $H_0 : \beta_1 = \gamma_1$ against $H_a : \beta_1 \neq \gamma_1$.

Problem 2. (A) Suppose women always married men who were exactly 5% plus 2 inches taller. Denote by Y and X the height, in inches, of the wife and husband, respectively. Determine the relation between X and Y and compute the correlation between X and Y.

(B) Compute the correlation coefficient for the following set of numbers (x, y):

$$(-2, 5),\ (-4, 4),\ (-6, 3),\ (-8, 2),\ (-10, 1).$$

Suggestion: draw a picture.

Problem 3. (a) The following results were obtained for about 1,000 families: average height of husband 68 inches, SD 2.5 inches; average height of wife 63 inches, SD 2.5 inches, correlation coefficient $r = 0.6$. Of the men who were married to women of height 60 inches, what percentage were under 64 inches? Assume normality wherever necessary.

(b) For the first-year students at a certain university, the correlation between SAT scores and first-year GPA was 0.60. Assume the distribution of the scores is jointly normal. Predict

the percentile rank on the first-year GPA for a student whose percentile rank on the SAT was (a) 90% (b) 30% (c) 50% (d) unknown

Problem 4. Let (X_i, Y_i), $i = 1, \dots, n$ be a random sample from a bivariate normal distribution. Confirm that (a) the sample correlation coefficient r is the MLE of the correlation coefficient ρ, and (b) if $\rho = 0$, then $r\sqrt{n-2}/\sqrt{1-r^2}$ has t_{n-2} distribution [here, you can relate the expression to the estimate of the slope of the regression line after conditioning on X; some ideas are in the book, Section 11.8].

This is a hard problem. In particular, the typical simplifying assumption that we know $\mathbb{E}X_i = \mathbb{E}Y_i = 0, \mathbb{E}X_i^2 = \mathbb{E}Y_i^2 = 1$ does not simplify the this problem but, in fact, makes it even harder: the MLE of the correlation coefficient is no longer r but instead a root of a rather complicated equation. A research paper on the subject is near the bottom of the class web page.

Problem 5. In the simple linear regression model $Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$, $i = 1, \dots, n$, assume that $x_i = x_{i,n} = i/n$ and ε_i are iid $\mathcal{N}(0, \sigma^2)$. Show that, as $n \to \infty$, the random vector $\sigma\sqrt{n}(\widehat{\beta}_0 - \beta_0, \widehat{\beta}_1 - \beta_1)^\top$ converges in distribution to a bivariate normal vector with zero mean and covariance matrix [written row-by-row] $(4, -6; -6, 12)$.

Homework 9.

Problem 1. Consider a collection of independent random variables

$$\eta_1, \dots, \eta_n, \varepsilon_1, \dots, \varepsilon_n, \zeta_1, \dots, \zeta_n,$$

each with mean zero, and assume that each ε_i and each ζ_i are normal with variance $\sigma^2 > 0$ and each η_i has variance $v^2 > 0$, but is not necessarily normal. Next, define $U_i = \mu_1 + \eta_i + \varepsilon_i$, $Y_i = \mu_2 + \eta_i + \zeta_i$. Confirm that (1) for each i, the random variables U_i and Y_i are not normal unless η_i is normal and are not independent: $\mathrm{Cov}(U_i, Y_i) = v^2$, whereas (2) the random variables $U_i - Y_i$, $i = 1, \dots, n$ are independent and normally distributed with mean $\mu_1 - \mu_2$ and variance $2\sigma^2$.

Problem 2. Let $\theta_1, \dots, \theta_n$ be real numbers. Confirm that $\theta_1 = \theta_2 = \ldots = \theta_n$ if and only if $a_1\theta_1 + \cdots + a_n\theta_n = 0$ for all real numbers $a_1, \dots, a_n$ satisfying $a_1 + \cdots + a_n = 0$. [This is obvious in one direction, in the other direction, consider several special collections of a_k with $a_k = 1, a_{k+1} = -1$ and all other $a_i = 0$.]

Problem 3. One of the ANOVA tools is variance stabilization.[2] Here is the main idea. Let Y be a random variable such that $\mathbb{E}Y = \theta$ and $\mathrm{Var}(Y) = f(\theta)$ for some function f.

(a) Let $g = g(y)$ be a smooth function. Using Taylor expansion $[g(Y) \approx g(\theta) + g'(\theta)(Y - \theta)]$, verify that $\mathrm{Var}\big(g(Y)\big) \approx \big(g'(\theta)\big)^2 f(\theta)$.

(b) If $g(y)$ is a constant multiple of an anti-derivative of $\big(f(y)\big)^{-1/2}$, then the above approximate variance does not depend on θ. This choice of g is called
`(approximately) variance-stabilizing transform`.

(c) Confirm that (approximately) variance-stabilizing transform for Poisson distribution $[f(\theta) = \theta]$ is $\sqrt{y}$ and for Binomial(n, p) distribution $[f(\theta) = \theta(1 - (\theta/n))]$, it is $\sin^{-1}\sqrt{y/n}$; $\sin^{-1}$ is the inverse sine. [This last one is unexpectedly tricky: you need to identify just the right way of integrating $\int dt/\sqrt{t(1-t)}$; all other ways would require some obscure trig identities to get the final answer]

(d) What will you get for the normal with mean and variance both equal to θ? [Same as for Poisson.] How about $\mathcal{N}(\theta, \theta^2)$? What can you do for the general normal distribution? [The previous approach will not work: in general normal distribution, mean and variance are not connected]

[2] A research paper on the subject is near the bottom of the class web page.

Problem 4. Consider one-way layout model in the form

$$Y_{ij} = \theta_i + \varepsilon_{ij}, \quad \varepsilon_{ij} \text{ are iid } \mathcal{N}(0, \sigma^2), \ i = 1, \dots, k, \ j = 1, \dots, n_i.$$

Show that, for every collection of real numbers $a_1, \dots, a_k$, the random variable $\sum_{i=1}^k a_i \bar{Y}_{i\bullet}$ is normal with mean $\sum_{i=1}^k a_i \theta_i$ and variance $\sigma^2 \sum_{i=1}^k a_i^2/n_i$.

Problem 5. In the setting of the previous problem, let $b_1, \dots, b_k$ be another collection of real numbers. Show that

$$\mathrm{Cov}\Big(\sum_{i=1}^k a_i \bar{Y}_{i\bullet}, \sum_{i=1}^k b_i \bar{Y}_{i\bullet} \Big) = \sigma^2 \sum_{i=1}^k \frac{a_i b_i}{n_i}.$$

Computations can be simplified if you convince yourself that

$$\mathrm{Cov}\big(\sum_i X_i, \sum_j Y_j \big) = \sum_{i,j} \mathrm{Cov}(X_i, Y_j)$$

Homework 10.

Problem 1. The table below presents the insurance rates, in dollars per six months, charged by different insurance companies I in different locations L for a similar product. Based on the numbers, will you conclude that the rates depend on the location? on the company?

$L \backslash I$	$I1$	$I2$	$I3$	$I4$	$I5$
$L1$	730	745	668	1065	1202
$L2$	836	725	618	869	1172
$L3$	1492	1384	1214	1502	1682
$L4$	996	884	802	1571	1272

Try two different approaches: (a) two separate one-ways layouts (one for I, the other for L). (b) randomized block design, considering I and L together. Then comment on the results.

Problem 2. Consider randomized block design model in the form

$$Y_{ij} = \theta_i + \beta_j + \varepsilon_{ij}, \quad \varepsilon_{ij} \text{ are iid } \mathcal{N}(0, \sigma^2), \ i = 1, \dots, k, \ j = 1, \dots, \ell.$$

Given real numbers $a_1, \dots, a_k$, and $b_1, \dots, b_\ell$, derive the distribution of the random variables $\sum_{i=1}^k a_i \bar{Y}_{i\bullet}$ and $\sum_{j=1}^\ell b_j \bar{Y}_{\bullet j}$.

Problem 3. Consider a two-factor model

$$Y_{ijl} = \theta_{ij} + \varepsilon_{ijl}, \ i = 1, \dots, K, \ j = 1, \dots, M, \ l = 1, \dots, L_{ij}, \ \varepsilon_{ijl} \text{ are iid } \mathcal{N}(0, \sigma^2).$$

Design a test of

$$H_0 : \theta_{ij} = \theta \text{ for all } i, j,$$

against the alternative that some of θ_{ij} are different.

Problem 4. List all 3×3 Latin squares with symbols A, B, C [there are 12 of those: two for each of the six permutations of (A, B, C)]. Convince yourself that if you restrict the first row AND column to (A, B, C), then there will be only one Latin square. If you still have time to spare, see if you can find the four different 4×4 Latin squares with (A, B, C, D) as the first row and column.

Problem 5. Convince yourself that, for every positive integers m, n and every $\alpha \in (0, 1)$,

$$t_{n,\alpha/2} \le \sqrt{m F_{m,n,\alpha}}.$$

Illustrate the result with a picture. What happens as $n \to \infty$? Here is an outline: (a) confirm that if $a > 0$ and $p > q$, then $P(\chi_p^2 > a) > P(\chi_q^2 > a)$; (b) note that $t_n^2 = \chi_1^2/(\chi_n^2/n)$, $P(t_n^2 > t_{n,\alpha/2}^2) = \alpha$, and $mF_{m,n} = \chi_m^2/(\chi_n^2/n)$, and the right-hand side stochastically dominates t_n^2 (because χ_m^2 stochastically dominates χ_1^2, $m > 1$).

Homework 11.

Problem 1. Here are the results of 100 rolls of a die:

Value	1	2	3	4	5	6
No. of times	24	12	12	11	11	30

Would you consider the die fair? Explain.

Problem 2. As part of a study on the selection of grand juries in Alameda county, the educational level of grand jurors was compared with the county distribution:

Educational level	County	Number of jurors
Elementary	28.4%	1
Secondary	48.5%	10
Some college	11.9%	16
College degree	11.2%	35
Total	100.0%	62

Could a simple random sample of 62 people from the county show a distribution of educational level so different from the county-wide one? Choose one option and explain.
(i) This is absolutely impossible.
(ii) This is possible, but fantastically unlikely.
(iii) This is possible but unlikely-the chance is around 1 % or so.
(iv) This is quite possible-the chance is around 10% or so.
(v) This is nearly certain.

Problem 3. In a certain town, there are about one million eligible voters. A simple random sample of size 10,000 was chosen, to study the relationship between sex and participation in the last election. The results:

	Men	Women
Voted	2,792	3,591
Didn't vote	1,486	2,131

Carry out a χ^2-test of the null hypothesis that sex and voting are independent and compute the p value.

Problem 4. In a company of 200 employees, there are 32 employees making at least $100,000 a year. There are 47 employees in the company that have a graduate degree. There are 143 employees that do not have a graduate degree and earn less than $100,000 per year. Based on these number, will you conclude that level of education and salary are dependent?

Problem 5. In a certain town, there are exactly 10,000 residents. The table below summarizes the relationship between sex and participation in the most recent election.

	Men	Women
Voted	2,825	3,575
Didn't vote	1,475	2,125

Are sex and voting participation independent? [Note: this problem is NOT about chi-square test.]

Homework 12.

Problem 1. A genetic model [a pretty famous one, known as the `Hardy-Weinberg equilibrium`] states that the proportion of offsprings in three classes should be $p^2, 2p(1-p)$, and $(1-p)^2$ for some $p \in (0,1)$ [note that this is just Binomial distribution $\mathcal{B}(2,p)$.]

An experiment yielded frequencies 30, 40, and 30 for the respective three classes.

(a) Does the model fit the data? [Start by computing the MLE of p].

(b) Do the data support the hypothesis that the model holds with $p = 1/2$?

(c) What is the different between the questions you are trying to answer in parts (a) and (b)?

Problem 2. Consider the following nine pairs (X_i, Y_i):

$$(9.4, 10.3),\ (7.8, 8.9),\ (5.6, 4.1),\ (12.1, 14.7),\ (6.9,\ 8.7),\ (4.2, 7.1),$$
$$(8.8, 11.3),\ (7.7, 5.2),\ (6.4,\ 7.8).$$

Assume that this is a random sample from two populations X and Y. The null hypothesis is that X and Y have the same distribution; the alternative is that they do not.

(a) Estimate the p-value of the sign test.

(b) Assuming that X and Y are normal with the same standard deviation, estimate the p-value of the t-test.

(c) Which p-value is bigger and does it make sense?

(d) Compute the sample correlation coefficient and the Spearman correlation coefficient for this sample, and comment on the results.

Problem 3. Assume that populations X and Y have continuous probability distributions. Convince yourself that, under the null hypothesis that X and Y have the same distribution,

$$\mathbb{E}(T^+) = \mathbb{E}(T^-) = \frac{n(n+1)}{4}$$

and

$$\mathrm{Var}(T^+) = \mathrm{Var}(T^-) = \frac{n(n+1)(2n+1)}{24},$$

where $T^{\pm}$ is the sum of ranks of positive/negative differences.

[Here is an idea: because $P(X_i = Y_i) = 0$, no ties are possible, and so $W = T^+ - T^- = \sum_{k=1}^n \varepsilon_k k$ is the total sum of signed ranks, where $\varepsilon_k = \pm 1$ are iid. Under the null hypothesis, $P(\varepsilon_k = 1) = 1/2$, and so $\mathbb{E}W = 0$, $\mathrm{Var}(W) = \sum_{k=1} k^2 = n(n+1)(2n+1)/6$. On the other hand, $T^+ + T^- = \sum_{k=1}^n k = n(n+1)/2$, so that $T^+ = \big(W + (n(n+1)/2)\big)/2$].

Problem 4. Assume that populations X and Y have continuous probability distributions. Let $x_1, \dots, x_m$ be a random sample from X and let $y_1, \dots, y_n$ be a random sample from Y. Let $u_{(1)}, \dots, u_{(m+n)}$ be the order statistics of the pooled sample $x_1, \dots, x_m, y_1, \dots, y_n$. Confirm that

$$U_x = mn + \frac{m(m+1)}{2} - W_x,$$

where W_x is the rank sum for the sample from X and U_x is the corresponding `Mann-Whitney statistic`, that is, the sum of the numbers of x-s that precede each of the y-s in the ordered list $u_{(1)}, \dots, u_{(m+n)}$. Then confirm that, under the null hypothesis that X and Y have the same distribution, we have $\mathbb{E}(U_x) = mn/2$ and $\mathrm{Var}(U_x) = mn(m+n+1)/12$.

[Here is an idea: with $N = m + n$, $W_x = \sum_{k=1}^N \varepsilon_k k$, where each ε_k takes value 0 or 1; since $\sum_{k=1}^N \varepsilon_k = m$, there is dependence. Under H_0, $P(\varepsilon_k = 1) = m/N$ and $P(\varepsilon_k = 1, \varepsilon_l = 1) = m(m-1)/(N(N-1))$. Then $\mathbb{E}(W_x) = (m/N)(N(N+1)/2)$, and the formula for $\mathbb{E}(U_x)$ follows. After rather long computations, the variance-covariance expansion leads to $\mathrm{Var}(W_x) = \frac{mn(N+1)}{12}$. Keep in mind that $\mathrm{Cov}(\varepsilon_k, \varepsilon_l) = -mn/((N^2(N-1))$ and $\sum_{l=1}^N l^3 = (N(N+1))^2/4$.]

Problem 5. A coin-making machine produces quarters in such a way that, for each coin, the probability p to turn up heads is uniform on $[0, 1]$. A coin pops out of the machine. Compute the conditional distribution, Bayesian point estimator, and a 95% credible interval for p given that the coin is

(a) Flipped once and lands heads;

(b) Flipped twice and lands heads once;

(c) Flipped three times and lands heads three times;

(d) Flipped 2000 times and lands heads 1500 times.

(e) Flipped N times and lands heads $n \le N$ times.

Now, repeat parts (a)–(e) under the assumption that there is some reason to believe that the coins from the machine are more likely to land heads so that the prior distribution for p is Beta with parameters 3 and 2 [so that the prior mean is 3/5].

Problem 6.

(1) Show that, for every fixed $x \in \mathbb{R}$, $\lim_{n\to\infty} \sqrt{n}\big(\hat{F}_n(x) - F(x)\big) = \mathcal{N}\big(0, F(x)(1-F(x))\big)$ in distribution. [Once you decipher the notations, this becomes a CLT result for Binomial distribution.]

(2) Confirm that if X and Y are independent and $F_Y(x) \le F_X(x)$ for all x, then $\mathbb{P}(Y \ge X) \ge 1/2$ [Note that $\mathbb{P}(Y \ge X) = 1 - \mathbb{E}F_Y(X)$ and $\mathbb{E}F_X(X) = 1/2$; you are welcome to assume that X and Y have pdf-s].

(3) If there are no ties, then

$$0 \le \sum_{i=1}^{n} \big(R(X_i) - R(Y_i)\big)^2 \le \frac{n(n^2-1)}{3}.$$

This is a particular case of a very famous result known as the rearrangement inequality: the largest value of $\sum R(X_i)R(Y_i)$ happens when $R(X_i) = R(Y_i)$, and the smallest, when $R(X_i) = n + 1 - R(Y_i)$. It also helps to note that

$$\sum_i R^2(X_i) = \sum_i R^2(Y_i) = 1^2 + 2^2 + 3^2 + \cdots + n^2 = \frac{n(n+1)(2n+1)}{6}.$$

(4) Consider a "multiplicative shift" model, in which the population X with pdf $f_X = f_X(x)$ and the population Y with pdf $f_Y = f_Y(x)$ are related by

$$f_Y(x) = \frac{1}{\theta} f_X(x/\theta),\ \theta > 0,$$

and the question is to determine whether $\theta = 1$. To simplify things further, assume that $f_X(x) = f_X(-x)$ for all x. Confirm that, by considering the random variables $\widetilde{X} = \ln|X|$, $\widetilde{Y} = \ln|Y|$, the problem is reduced to the standard shift model.