Practice Final Exam

Math 170S (Winter 2025)
Instructor: Koffi Enakoutsa
TA: Colin Ni

Name: ____

Student ID: ____

Show all work. You may use one sheet (front and back) of notes. You may not use a calculator. Please ensure your phone is silenced and stowed away.

Duration: 3 hours.

The following is my own work, without the aid of any other person.

Signature: ____

Problem 1

Consider the following 20 sorted samples $X_1, \dots, X_{20}$ drawn from the random variable $X \sim \mathrm{Norm}(1, 2)$:

−0.1	0.13	0.6	0.77	1.28	1.39	1.51	1.57	1.65	1.74
2.04	2.04	2.06	2.15	2.53	2.6	2.73	3.05	3.23	3.69.

The sum of the samples and the sum of the squares of the samples are respectively

$$\sum_{i=1}^{20} X_i = 36.66 \qquad \text{and} \qquad \sum_{i=1}^{20} X_i^2 = 86.20.$$

(*i*) (5 points) Show that

$$\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \bar{X}^2.$$

(*ii*) (5 points) Use part (*i*) to compute the sample mean $\bar{X}$ and the sample variance s_X^2.

(*iii*) (10 points) Compute the third quartile $\pi_{0.75}$.

(*iv*) (5 points, challenge) More generally, let $X_1, \dots, X_n$ be samples drawn from a random variable X with variance $\mathrm{Var}(X) = \sigma^2$. The sample variance s_X^2 is an estimator for σ^2, but so is

$$v_X := \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2.$$

In one or two sentences, explain the key difference between these two estimators and how it relates to Bessel's correction.

Solution. For (*i*), we have

$$\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \frac{1}{n}\sum_{i=1}^{n} 2X_i\bar{X} + \frac{1}{n}\sum_{i=1}^{n}\bar{X}^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2 - 2\bar{X}^2 + \bar{X}^2$$

For (*ii*), the sample mean is of course

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{20} X_i = \frac{36.66}{20}.$$

To compute the sample variance, observe that

$$s_X^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2 = \frac{n}{n-1}\left(\sum_{i=1}^{n} X_i^2 - \bar{X}^2\right),$$

so here

$$s_X^2 = \frac{20}{19}\left(86.20 - \left(\frac{36.66}{20}\right)^2\right).$$

For (*iii*), observe that 0.75 is between 15/21 and 16/21. In fact

$$0.75 = (1-t)\frac{15}{21} + t\frac{16}{21}$$

for $t = 1/4$. Thus

$$\pi_{0.75} = (1 - 1/4)X_{(15)} + (1/4)X_{(16)} = \frac{3}{4}\cdot 2.53 + \frac{1}{4}\cdot 2.6.$$

For (*iv*), sample variance s_X^2 is an unbiased estimator for σ^2, whereas v_X is biased. Bessel's correction is the correction $s_X^2 = \frac{n}{n-1}v_X^2$ that makes s_X^2 unbiased. □

Problem 2

In this problem, we will explore the order statistics of the so-called power distribution. The *power distribution* $P(\alpha)$ *with power* $\alpha > 0$ is defined by the pdf

$$f_{P(\alpha)}(x) = \alpha x^{\alpha-1} \qquad \text{for } 0 \leq x \leq 1.$$

Let $P_1, \ldots, P_n$ be samples from $P(\alpha)$.

(i) (5 points) Check that $f_{P(\alpha)}$ is a valid pdf.

(ii) (10 points) Determine the distribution of the nth order statistic $P_{(n)}$ in terms of a power distribution.

(iii) (10 points) The *Kumaraswamy distribution* $K(a, b)$ *with shape* $a, b \geq 0$ is defined by the pdf

$$f_{K(a,b)}(x) = abx^{a-1}(1 - x^a)^{b-1} \qquad \text{for } 0 \leq x \leq 1.$$

(Note the similarity with $\mathrm{Beta}(\alpha, \beta)$ which has pdf proportional to $x^{\alpha-1}(1-x)^{\beta-1}$.) Determine the distribution of the 1st order statistic $P_{(1)}$ in terms of a Kumaraswamy distribution.

Solution. For *(i)*, we simply need to check that

$$\int_0^1 f_{P(\alpha)}(x)\,\mathrm{d}x = \int_0^1 \alpha x^{\alpha-1}\,\mathrm{d}x = x^\alpha\Big|_0^1 = 1.$$

For *(ii)*, let us use the general formula

$$f_{X_{(r)}}(x) = r\binom{n}{r} F_X(x)^{r-1}(1 - F_X(x))^{n-r} f_X(x)$$

for the rth order statistic of a random variable X from week 1 discussion section. Since $F_{P(\alpha)}(x) = x^\alpha$, we have

$$f_{P_{(n)}}(x) = nx^{\alpha(r-1)} \cdot \alpha x^{\alpha-1} = n\alpha x^{n\alpha-1} = f_{P(n\alpha)}(x).$$

Thus the nth order statistic $P_{(n)} \sim P(n\alpha)$ follows the power distribution with power $n\alpha$.

For *(iii)*, by the same general formula, we have

$$f_{P_{(1)}}(x) = \alpha n x^{\alpha-1}(1 - x^\alpha)^{n-1} = f_{K(\alpha,n)}(x).$$

Thus the 1st order statistic $P_{(1)} \sim K(\alpha, n)$ follows the Kumaraswamy distribution with shape α, n. $\square$

Problem 3

In this problem, we will find the MLE for the probability of success for a geometric random variable. Recall that a geometric random variable Geom(p) with probability of success $0 \leq p \leq 1$ has pdf

$$f_{\text{Geom}(p)}(x) = (1-p)^{x-1}p \qquad \text{for } x \geq 0.$$

(i) (5 points) Describe precisely the likelihood function.

(ii) (15 points) Find the maximum likelihood estimator $\hat{p}$ of the parameter p.

(iii) (5 points) Explain why the estimator $\hat{p}$ you found makes sense intuitively in terms of what a geometric random variable models and what p represents.

Solution. For *(i)*, the likelihood of a sample $X_1, \ldots, X_n$ given a value of p is

$$\begin{aligned} L(X_1, \ldots, X_n \mid p) &= f_{\text{Geom}(p)}(X_1) \cdots f_{\text{Geom}(p)}(X_n) \\ &= (1-p)^{X_1 - 1}p \cdots (1-p)^{X_n - 1}p \\ &= (1-p)^{n\bar{X} - n}p^n. \end{aligned}$$

For *(ii)*, maximizing the likelihood is equivalent to maximizing the log-likelihood

$$\ln L(X_1, \ldots, X_n \mid p) = (n\bar{X} - n)\ln(1-p) + n\ln(p).$$

The critical points of of the log-likelihood occur where

$$0 = \frac{\mathrm{d}\ln L}{\mathrm{d}p}(X_1, \ldots, X_n \mid p) = -\frac{n\bar{X} - n}{1-p} + \frac{n}{p}.$$

Solving for p shows the log-likelihood has a unique critical point $\hat{p} = 1/\bar{X}$. To see that this is a maximum, let us compute the second derivative:

$$\frac{\mathrm{d}^2 \ln L}{\mathrm{d}p^2}(X_1, \ldots, X_n \mid p) = -\frac{n\bar{X} - n}{(1-p)^2} - \frac{n}{p^2}.$$

In order to plug in $\hat{p} = 1/\bar{X}$, we must handle the case where $1/\bar{X} = 1$ separately. In this case the log-likelihood is simply $n\ln(p)$, so since $0 \leq p \leq 1$, the value $p = 1 = 1/\bar{X}$ is the unique maximizing value. Otherwise $1/\bar{X} \neq 1$, in which case

$$\frac{\mathrm{d}^2 \ln L}{\mathrm{d}p^2}\left(X_1, \ldots, X_n \;\middle|\; \frac{1}{\bar{X}}\right) = -\frac{n\bar{X} - n}{(1 - \frac{1}{\bar{X}})^2} - \frac{n}{\frac{1}{\bar{X}}^2} < 0.$$

Therefore, the MLE $\widehat{p}$ of the parameter p for the geometric distribution Geom(p) is given by

$$\widehat{p} = \frac{1}{\bar{X}}.$$

For *(iii)*, this makes sense intuitively because $\bar{X}$ is the average number of trials needed to get a success, so $\frac{1}{\bar{X}}$ estimates the probability of success. $\square$

Problem 4

In this problem, we will find the method of moments estimators for the parameters of a Gamma distribution. Recall that a Gamma distribution Gamma(α, θ) has parameters $\alpha > 0$ and $\theta > 0$ and is defined by the pdf

$$f_{\Gamma(\alpha,\theta)}(x) = \frac{1}{\Gamma(\alpha)\theta^\alpha} x^{\alpha-1} e^{-\frac{x}{\theta}} \qquad \text{for } x \geq 0.$$

Recall further that it has mean $\alpha\theta$ and variance $\alpha\theta^2$.

(*i*) (10 points) Let $X \sim \text{Gamma}(\alpha, \theta)$. Find the method of moments estimators $\tilde{\alpha}$ and $\tilde{\theta}$ for the parameters α and θ respectively.

(*ii*) (5 points) Set

$$v_X^2 = \frac{n-1}{n} s_X^2 = \frac{1}{n} \sum_{i=1}^n X_i^2 - \bar{X}^2.$$

Write $\tilde{\alpha}$ and $\tilde{\theta}$ in terms of V and $\bar{X}$.

(*ii*) (10 points) Method of moments estimators can sometimes produce nonsensical estimates. Elaborate on this point, and provide a concrete example using the estimators $\tilde{\alpha}$ and $\tilde{\theta}$.

Solution. For (*i*), we are given that the theoretical first moment is $\mathbb{E}(X) = \alpha\theta$, and we deduce that the theoretical second moment is

$$\mathbb{E}(X^2) = \text{Var}(X) + \mathbb{E}(X)^2 = \alpha\theta^2 + \alpha^2\theta^2 = \alpha\theta^2(1 + \alpha).$$

Suppose our sample is $X_1, \ldots, X_n$. Our estimators for the first and second moments are respectively

$$\hat{M}_1 = \frac{1}{n} \sum_{i=1}^n X_i \qquad \text{and} \qquad \hat{M}_2 = \frac{1}{n} \sum_{i=1}^n X_i^2.$$

Thus, the method of moments estimators $\tilde{\alpha}$ and $\tilde{\theta}$ are the solution to the system of equations

$$\tilde{\alpha}\tilde{\theta} = \hat{M}_1 \qquad \text{and} \qquad \tilde{\alpha}\tilde{\theta}^2(1 + \tilde{\alpha}) = \hat{M}_2.$$

Some algebra shows that

$$\tilde{\alpha} = \frac{\hat{M}_1}{\hat{M}_2 - \hat{M}_1^{\,2}} \qquad \text{and} \qquad \tilde{\theta} = \frac{\hat{M}_2 - \hat{M}_1^{\,2}}{\hat{M}_1}.$$

For (*ii*), simply write

$$\tilde{\alpha} = \frac{\bar{X}}{V} \qquad \text{and} \qquad \tilde{\theta} = \frac{V}{\bar{X}}.$$

For (*iii*), for this example there are many ways these estimators can give nonsensical answers. For example, α and θ are always positive real numbers, but the estimator $\tilde{\alpha}$ can be zero when $\bar{X} = 0$ (this can happen because the support is $x \geq 0$) and the estimator $\tilde{\theta}$ can be zero when $V = 0$. Even worse, when $\bar{X} = 0$, the estimator $\tilde{\theta}$ is undefined, and when $V = 0$, the estimator $\tilde{\alpha}$ is undefined. $\square$

Problem 5

In this problem, we will work out an example of linear regression and discuss what happens when we introduce regularization. Consider the following dataset of x-values $x_1, \ldots, x_8$ and y-values $y_1, \ldots, y_8$:

$$\begin{array}{lcccccccc} x: & -2 & 4 & -1 & -2 & 0 & 3 & 1 & 1 \\ y: & 4 & -7 & 3 & 5 & 2 & -4 & -1 & -2. \end{array}$$

(*i*) (5 points) Plot this dataset.

(*ii*) (15 points) Find the exact line of best fit using linear regression. (*Hint*: Your answer should contain the number 32/17.)

(*iii*) (5 points, challenge) Recall that for the linear model $y(x) = \alpha + x\beta$, linear regression finds the parameters α and β that minimizes the mean squared error. Consider the regularized mean squared error function

$$\frac{1}{n}\sum_{i=1}^{n}(y(x_i) - y_i)^2 + \frac{\lambda}{n}(\alpha^2 + \beta^2)$$

where $\lambda > 0$ is some hyperparameter. This penalizes α and β being large, resulting in a simpler model. Find α and β that minimize this regularized error function. (*Hint*: Write the function as $\frac{1}{n}(\mathbf{X}\gamma - \mathbf{y})^T(\mathbf{X}\gamma - \mathbf{y}) + \frac{\lambda}{n}\gamma^T\gamma$ which has derivative $\frac{2}{n}(\mathbf{X}^T\mathbf{X}\gamma - \mathbf{X}^T\mathbf{y}) + \frac{2\lambda}{n}\gamma$.)

Solution. For (*i*), see this Wolfram Alpha query.

For (*ii*), see week 5 discussion section. We have

$$\mathbf{X} = \begin{pmatrix} 1 & -2 \\ 1 & 4 \\ 1 & -1 \\ 1 & -2 \\ 1 & 0 \\ 1 & 3 \\ 1 & 1 \\ 1 & 1 \end{pmatrix} \quad \text{and} \quad \mathbf{y} = \begin{pmatrix} 4 \\ -7 \\ 3 \\ 5 \\ 2 \\ -4 \\ -1 \\ -2 \end{pmatrix}.$$

So

$$\mathbf{X}^T\mathbf{X} = \begin{pmatrix} 8 & 4 \\ 4 & 36 \end{pmatrix} \quad \text{and} \quad \mathbf{X}^T\mathbf{y} = \begin{pmatrix} 0 \\ -64 \end{pmatrix}.$$

Thus

$$\begin{pmatrix} \hat{\alpha} \\ \hat{\beta} \end{pmatrix} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} = \frac{1}{272}\begin{pmatrix} 36 & -4 \\ -4 & 8 \end{pmatrix}\begin{pmatrix} 0 \\ -64 \end{pmatrix} = \frac{1}{272}\begin{pmatrix} 256 \\ -512 \end{pmatrix}.$$

Therefore

$$y(x) = \frac{16}{17} - \frac{32}{17}x$$

is the line of best fit.

For (*iii*), setting the derivative equal to zero and solving for γ gives

$$\hat{\gamma} = (\mathbf{X}^T\mathbf{X} + \lambda I)^{-1}\mathbf{X}^T\mathbf{y}.$$

Unraveling gives $\hat{\alpha}$ and $\hat{\beta}$. Note that we have the theoretical benefit of $\mathbf{X}^T\mathbf{X} + \lambda I$ being guaranteed to be invertible, because $\lambda > 0$ makes this positive definite. $\square$

Problem 6

In this problem, we will give an example of a sufficient statistic and an example of a non-sufficient statistic. Let $X \sim \text{Pois}(\lambda)$ be an Poisson random variable with unknown rate $\lambda > 0$. Recall that the pdf is given by

$$f_{\text{Pois}(\lambda)}(k) = \frac{\lambda^k e^{-k}}{k!} \qquad \text{for } k \geq 0 \text{ an integer}$$

and that the MLE of the parameter λ is the sample mean: $\hat{\lambda} = \bar{X}$.

(*i*) (20 points) Show that the MLE $\hat{\lambda}$ is a sufficient statistic for λ.

(*ii*) (5 points, challenge) Show that the sample minimum is not a sufficient statistic for λ.

Solution. For (i), the joint pdf of a sample $X_1, \ldots, X_n$ and a rate $\lambda > 0$ is

$$f_{\text{Pois}(\lambda)}(X_1) \cdots f_{\text{Pois}(\lambda)}(X_n) = \frac{\lambda^{X_1} e^{-X_1}}{X_1!} \cdots \frac{\lambda^{X_n} e^{-X_n}}{X_n!} = \frac{\lambda^{X_1 + \cdots + X_n} e^{-(X_1 + \cdots + X_n)}}{X_1! \cdots X_n!}.$$

Since $X_1 + \cdots + X_n = n\bar{X} = n\hat{\lambda}$, we can factor this joint pdf as

$$\lambda^{n\hat{\lambda}} \cdot \frac{e^{-(X_1 + \cdots + X_n)}}{X_1! \cdots X_n!},$$

where the factor on the LHS only depends on the MLE $\hat{\lambda}$ and the parameter λ, and the RHS only depends on the sample $X_1, \ldots, X_n$. By our definition of sufficient statistics, this shows that $\hat{\lambda}$ is a sufficient statistic for λ.

For (ii), the intuition is that the sample minimum does not capture the same information as the sample mean. For example, the samples $0, 5$ and $0, 6$ have the same sample minimums but different means. More precisely, suppose for the sake of contradiction that for every n, the joint pdf can be factored as

$$\frac{\lambda^{X_1 + \cdots + X_n} e^{-(X_1 + \cdots + X_n)}}{X_1! \cdots X_n!} = \phi(X_{(1)}, \lambda) h(X_1, \ldots, X_n)$$

for some functions ϕ and h. Let us absorb the $X_1! \cdots X_n!$ into the function h by defining $H(X_1, \ldots, X_n) = h(X_1, \ldots, X_n) X_1! \cdots X_n!$ so that

$$\lambda^{X_1 + \cdots + X_n} e^{-(X_1 + \cdots + X_n)} = \phi(X_{(1)}, \lambda) H(X_1, \ldots, X_n).$$

Consider the case $n = 2$ and even more specifically when $X_1 = 0$:

$$\lambda^{X_2} e^{-X_2} = \phi(0, \lambda) H(0, X_2).$$

For $\lambda = 1$, we get $H(0, X_2) \propto e^{-X_2}$, and for $\lambda = 2$, we get $H(0, X_2) \propto 2^{X_2} e^{-X_2}$. This is a contradiction because e^{-X^2} and $2^{X_2} e^{-X_2}$ are not proportional. $\square$

Problem 7

In this problem, we will walk you through how to determine the Bayes estimate of the mean of a normal distribution with a normal prior, and then we will discuss how to interpret the Bayes estimate.

Let $X_1, \ldots, X_n$ be a sample from the normal random variable $\mathrm{Norm}(\mu, \sigma^2)$ where σ^2 is known, and assume that the prior distribution of μ is the normal distribution $\mathrm{Norm}(\mu_0, \sigma_0^2)$.

(*i*) (5 points) Recall that the pdf of $\mathrm{Norm}(\mu, \sigma^2)$ is

$$f_{\mathrm{Norm}(\mu,\sigma^2)}(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right).$$

Find a, b, c such that the joint distribution of $\bar{X}$ and μ is proportional to $\exp\left(\frac{a\mu^2 - b\mu}{c}\right)$.

(*ii*) (5 points) Observe that

$$\exp\left(\frac{a\mu^2 - b\mu}{c}\right) = \exp\left(\frac{(\mu - \frac{b}{2a})^2 - \frac{b^2}{4a^2}}{\frac{c}{a}}\right) \propto \exp\left(\frac{(\mu - \frac{b}{2a})^2}{\frac{c}{a}}\right).$$

Find the normal distribution whose pdf this is proportional to.

(*iii*) (5 points) Deduce that the posterior distribution of μ is

$$\mathrm{Norm}\left(\frac{\bar{X}\sigma_0^2 + \mu_0\sigma^2/n}{\sigma_0^2 + \sigma^2/n}, \frac{\sigma^2\sigma_0^2/n}{\sigma_0^2 + \sigma^2/n}\right).$$

(*iv*) (5 points) Let us take our Bayes estimate to be the mean of the posterior. Use this Bayes estimate to illustrate the idea that Bayes estimators are influenced by both the data and the prior.

(*v*) (5 points) Explain what it means for n to be large, and explain what happens to the Bayes estimate in this case. Similarly, explain what it means for σ_0^2 to be small, and explain what happens to the Bayes estimate in this case.

Solution. For (*i*), the joint distribution of $\bar{X} \sim \mathrm{Norm}(\mu, \sigma^2/n)$ and μ is

$$\begin{aligned}
&\frac{1}{\sqrt{2\pi\sigma^2/n}} \exp\left(-\frac{(\bar{X}-\mu)^2}{2\sigma^2/n}\right) \frac{1}{\sqrt{2\pi\sigma_0^2}} \exp\left(-\frac{(\mu-\mu_0)^2}{2\sigma_0^2}\right) \\
&\quad\propto \exp\left(-\frac{(\bar{X}-\mu)^2}{2\sigma^2/n} + \frac{(\mu-\mu_0)^2}{2\sigma_0^2}\right) \\
&\quad= \exp\left(-\frac{\sigma_0^2(\bar{X}-\mu)^2 + (\sigma^2/n)(\mu-\mu_0)^2}{2\sigma^2\sigma_0^2/n}\right) \\
&\quad= \exp\left(-\frac{(\sigma_0^2 - \sigma^2/n)\mu^2 + (2\sigma_0^2\bar{X} + 2\sigma^2\mu_0/n)\mu + (\sigma_0^2\bar{X}^2 + \sigma^2\mu_0^2/n)}{2\sigma^2\sigma_0^2/n}\right) \\
&\quad\propto \exp\left(-\frac{(\sigma_0^2 - \sigma^2/n)\mu^2 + (2\sigma_0^2\bar{X} + 2\sigma^2\mu_0/n)\mu}{2\sigma^2\sigma_0^2/n}\right).
\end{aligned}$$

Thus, take

$$a = \sigma_0^2 + \sigma^2/n, \qquad b = 2(\sigma_0^2\bar{X} + \sigma^2\mu_0/n), \qquad c = 2\sigma^2\sigma_0^2/n.$$

For (*ii*), this is proportional to the pdf of the normal distribution

$$\mathrm{Norm}\left(\frac{b}{2a}, \frac{c}{2a}\right).$$

For (iii), by Bayes' theorem, the posterior distribution is proportional to the joint distribution of the parameter and the prior. Thus by (i) and (ii) combined, this is proportional to the normal distribution described in (ii). Plugging in our a, b, c gives the result.

For (iv), we can write our Bayes estimate as a linear combination

$$\frac{\bar{X}\sigma_0^2 + \mu_0\sigma^2/n}{\sigma_0^2 + \sigma^2/n} = \frac{\sigma_0^2}{\sigma_0^2 + \sigma^2/n}\bar{X} + \frac{\sigma^2/n}{\sigma_0^2 + \sigma^2/n}\mu_0$$

of the MLE $\bar{X}$, which only sees the data, and the prior estimate μ_0.

For (v), of course n being large means that our sample size is large, and in this case the Bayes estimate becomes (in the limit) the MLE $\bar{X}$ and entirely ignores the prior. On the other hand, σ_0^2 being small means that our prior is strong (we strongly believe that μ should be close to μ_0), and in this case the Bayes estimate becomes (in the limit) the prior estimate μ_0. $\square$

Problem 8

In this problem, we will construct a confidence interval for the difference between two means assuming unknown but proportional variances. Assume $X \sim \mathrm{Norm}(\mu_X, \sigma^2)$ and $Y \sim \mathrm{Norm}(\mu_Y, \sigma^2)$ are normal, where the means are unknown and where the variances are equal but otherwise unknown. Suppose we collect a sample of size n_X from X with sample mean $\bar{X}$ and sample variance s_Y^2, and suppose we collect a sample of size n_Y from Y with sample mean $\bar{Y}$ and sample variance s_Y^2.

(*i*) (10 points) Show that

$$Z := \frac{(\bar{X} - \bar{Y}) - (\mu_X - \mu_Y)}{\sigma\sqrt{1/n_X + 1/n_Y}} \sim \mathrm{Norm}(0, 1).$$

(*ii*) (5 points, challenge) Show that

$$V := \frac{(n_X - 1)s_X^2}{\sigma^2} + \frac{(n_Y - 1)s_Y^2}{\sigma^2} \sim \chi^2(n_X + n_Y - 2).$$

(*iii*) (10 points) You may use the (kind of deep) fact that the Z and V from parts (*i*) and (*ii*) are independent. Show that

$$\left((\bar{X} - \bar{Y}) - t_0 s_P\sqrt{\frac{1}{n_X} + \frac{1}{n_Y}}, (\bar{X} - \bar{Y}) + t_0 s_P\sqrt{\frac{1}{n_X} + \frac{1}{n_Y}}\right)$$

where

$$s_P = \sqrt{\frac{(n_X - 1)s_X^2 + (n_Y - 1)s_Y^2}{n_X + n_Y - 2}}$$

and

$$t_0 = t_{\alpha/2}(n_X + n_Y - 2)$$

is a confidence interval for $\mu_X - \mu_Y$ with confidence coefficient $1 - \alpha$.

Solution. For (*i*), since affine linear combinations of normal random variables are normal, certainly Z is normal. Let us show that it is unit normal. Its mean is zero because

$$\mathbb{E}((\bar{X} - \bar{Y}) - (\mu_X - \mu_Y)) = (\mathbb{E}(\bar{X}) - \mu_X) - (\mathbb{E}(\bar{Y}) - \mu_Y) = 0,$$

where $\mathbb{E}(\bar{X}) = \mu_X$ since sample mean is an unbiased estimator of the mean, and similarly with Y. Its variance is one because

$$\mathrm{Var}((\bar{X} - \bar{Y}) - (\mu_X - \mu_Y)) = \mathrm{Var}(\bar{X} - \bar{Y}) = \mathrm{Var}(\bar{X}) + \mathrm{Var}(\bar{Y}) = \frac{\sigma^2}{n_X} + \frac{\sigma^2}{n_Y}$$

where the middle equality is by independence and the last equality is by a computation from discussion section week 2.

For (*ii*), recall that the chi-squared distribution $\chi^2(n)$ with n degrees of freedom can be defined as the distribution of a sum of squares of n independent unit normals. Observe that by defining $Z_i = (X_i - \mu)/\sigma \sim \mathrm{Norm}(0, 1)$, we have

$$\frac{(n_X - 1)s_X^2}{\sigma^2} = \sum_{i=1}^{n_X}\left(\frac{X_i - \bar{X}}{\sigma}\right)^2 = \sum_{i=1}^{n_X}(Z_i - \bar{Z})^2 \sim \chi^2(n_X - 1)$$

by the key property of chi-squared distributions from discussion section week 8. Similarly $(n_Y - 1)s_Y^2/\sigma^2 \sim \chi^2(n_Y - 1)$, so by independence their sum V follows the distribution $\chi^2(n_X + n_Y - 2)$.

For (iii), recall that the t-distribution $t(n)$ with n degrees of freedom can be defined as $Z/\sqrt{V/n}$, where $Z \sim \mathrm{Norm}(0,1)$ and $V \sim \chi^2(n)$ are independent. Thus we have

$$\frac{Z}{\sqrt{V/(n_X+n_Y-2)}} \sim t(n_X+n_Y-2).$$

Of course, a two-sided confidence interval for this quantity with confidence coefficient $1-\alpha$ is $(-t_0, t_0)$ where the definition $t_0 = t_{\alpha/2}(n_X+n_Y-2)$ is given in the problem. Thus, a two-sided confidence interval for $\mu_X - \mu_Y$ is

$$\left((\bar{X}-\bar{Y}) - t_0\sqrt{\frac{V}{n_X+n_Y-2}}\sigma\sqrt{\frac{1}{n_X}+\frac{1}{n_Y}}, \quad (\bar{X}-\bar{Y}) + t_0\sqrt{\frac{V}{n_X+n_Y-2}}\sigma\sqrt{\frac{1}{n_X}+\frac{1}{n_Y}}\right),$$

and it is easy to check that this is the same as

$$\left((\bar{X}-\bar{Y}) - t_0 s_P\sqrt{\frac{1}{n_X}+\frac{1}{n_Y}}, \quad (\bar{X}-\bar{Y}) + t_0 s_P\sqrt{\frac{1}{n_X}+\frac{1}{n_Y}}\right).$$

□

Problem 9

In this problem, we will test the hypothesis that the air quality in San Francisco is different from the air quality in Los Angeles. Let X and Y denote the concentration of suspended particles in the air in San Francisco and Los Angeles, respectively. As in Problem 9, assume $X \sim \text{Norm}(\mu_X, \sigma^2)$ and $Y \sim \text{Norm}(\mu_Y, \sigma^2)$ are normal, where the means are unknown and where the variances are equal but otherwise unknown. Suppose we collect data with the following statistics:

$$n_X = 5 \qquad \bar{X} = 79 \qquad s_X = 5$$
$$n_Y = 10 \qquad \bar{Y} = 84 \qquad s_Y = 5.$$

(*i*) (20 points) Using the confidence interval described in Problem 8 part (*iii*), perform a hypothesis test on whether $\mu_X \neq \mu_Y$ at confidence level $\alpha = 0.05$. You should refer to the table of normal and t-distribution values on the last page of this exam, and you may use that $\sqrt{3/10} \approx 11/20$

(*ii*) (5 points) Give a bound (as tight as possible using the table at the end of the exam) for the p-value of your test. You may use that $1/11 \approx 0.09$.

Solution. For (*i*), our null hypothesis is that $\mu_X = \mu_Y$, and our alternative hypothesis is that $\mu_X \neq \mu_Y$. Using Problem 8 part (*iii*), a confidence interval centered at $\bar{X} - \bar{Y}$ with confidence coefficient $1 - \alpha$ is given by

$$\left((\bar{X} - \bar{Y}) - t_0 s_P \sqrt{\frac{1}{n_X} + \frac{1}{n_Y}}, \qquad (\bar{X} - \bar{Y}) + t_0 s_P \sqrt{\frac{1}{n_X} + \frac{1}{n_Y}} \right),$$

so we can take $\bar{X} - \bar{Y}$ to be our test statistic and our critical region to be the region outside of the interval

$$\left(-t_0 s_P \sqrt{\frac{1}{n_X} + \frac{1}{n_Y}}, \qquad t_0 s_P \sqrt{\frac{1}{n_X} + \frac{1}{n_Y}} \right).$$

Plugging in everything gives

$$t_0 = t_{0.025}(13) = 2.160, \qquad s_P = 5, \qquad \sqrt{\frac{1}{5} + \frac{1}{10}} = \sqrt{\frac{3}{10}} \approx \frac{11}{20},$$

so our interval is

$$\left(-2.160 \cdot \frac{11}{4}, \qquad 2.160 \cdot \frac{11}{4} \right).$$

Since our test statistic $\bar{X} - \bar{Y} = -5$ is not in the critical region, we do not reject the null hypothesis.

For (*ii*), our t-score is

$$\frac{\bar{X} - \bar{Y}}{s_P \sqrt{\frac{1}{n_X} + \frac{1}{n_Y}}} \approx -\frac{20}{11} \approx 1.82,$$

which is between $t_{0.05}(13) \approx 1.771$ and $t_{0.025}(13) \approx 2.160$. Thus the p-value is between 0.05 and 0.10. $\square$

Table of normal and t-distribution values

The following table gives some common values of the cdf of some t-distributions. Since $t(\infty) = \mathrm{Norm}(0,1)$ this also contains common values of the cdf of the unit normal distribution. Recall that $t_\alpha(r)$ is defined as $t_\alpha(r) = F_{t(r)}(1-\alpha)$, where $t(r)$ denotes the t-distribution with r degrees of freedom.

r	$t_{0.10}(r)$	$t_{0.05}(r)$	$t_{0.025}(r)$	$t_{0.01}(r)$	$t_{0.005}(r)$	$t_{0.001}(r)$
1	3.078	6.314	12.706	31.821	63.657	318.31
2	1.886	2.920	4.303	6.965	9.925	22.327
3	1.638	2.353	3.182	4.541	5.841	10.215
4	1.533	2.132	2.776	3.747	4.604	7.173
5	1.476	2.015	2.571	3.365	4.032	5.893
6	1.440	1.943	2.447	3.143	3.707	5.208
7	1.415	1.895	2.365	2.998	3.499	4.785
8	1.397	1.860	2.306	2.896	3.355	4.501
9	1.383	1.833	2.262	2.821	3.250	4.297
10	1.372	1.812	2.228	2.764	3.169	4.144
11	1.363	1.796	2.201	2.718	3.106	4.025
12	1.356	1.782	2.179	2.681	3.055	3.930
13	1.350	1.771	2.160	2.650	3.012	3.852
14	1.345	1.761	2.145	2.624	2.977	3.787
15	1.341	1.753	2.131	2.602	2.947	3.733
16	1.337	1.746	2.120	2.583	2.921	3.686
17	1.333	1.740	2.110	2.567	2.898	3.646
18	1.330	1.734	2.101	2.552	2.878	3.610
19	1.328	1.729	2.093	2.539	2.861	3.579
20	1.325	1.725	2.086	2.528	2.845	3.552
21	1.323	1.721	2.080	2.518	2.831	3.527
22	1.321	1.717	2.074	2.508	2.819	3.505
23	1.319	1.714	2.069	2.500	2.807	3.485
24	1.318	1.711	2.064	2.492	2.797	3.467
25	1.316	1.708	2.060	2.485	2.787	3.450
26	1.315	1.706	2.056	2.479	2.779	3.435
27	1.314	1.703	2.052	2.473	2.771	3.421
28	1.313	1.701	2.048	2.467	2.763	3.408
29	1.311	1.699	2.045	2.462	2.756	3.396
30	1.310	1.697	2.042	2.457	2.750	3.385
35	1.306	1.690	2.030	2.438	2.724	3.340
40	1.303	1.684	2.021	2.423	2.704	3.307
45	1.301	1.679	2.014	2.412	2.690	3.281
50	1.299	1.676	2.009	2.403	2.678	3.261
55	1.297	1.673	2.004	2.396	2.668	3.245
60	1.296	1.671	2.000	2.390	2.660	3.232
∞	1.282	1.645	1.960	2.326	2.576	3.090