1 Problem 1: Probability

Consider a random variable X with cumulative distribution function (CDF)

$$F(x) = \begin{cases} 1 - e^{-x^2/2}, & x \geq 0, \\ 0, & \text{otherwise.} \end{cases}$$

(a) Find the density function of X.

(b) Derive the mean of X. *Hint: integration by parts*

(c) Consider a random variable $U \sim \text{Uniform}(0,1)$, i.e., a random variable that is uniformly distributed on the interval $(0,1)$, show that $Y = \sqrt{-2\log U}$ has the same distribution as X.

2 Problem 2: T-test and chi-square distribution

A biologist wishes to figure out the mean of the length of the petals of a flower. The person who has conducted the study assumes that the measurements follow $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mu, \sigma^2)$, with unknown parameters $\mu \in \mathbb{R}_{>0}$ and $\sigma > 0$. Unfortunately, they have dropped the original data and have only recorded the three numbers $\sum_{i=1}^n X_i = 34$, $\sum_{i=1}^n X_i^2 = 65$, and $n = 25$. *Hint: quantiles of the t-distribution and chi-square distribution can be found at the end of the document.*

(a) Based on this summary data, would it be possible for them to conduct a t-test for $H_0 : \mu = 1$ versus $H_1 : \mu \neq 1$? If no, explain your reasoning. If it is possible, conduct the test at level $\alpha = .05$.

(b) Based on this summary data, conduct an appropriate test for $H_0 : \sigma = .5$ versus $H_1 : \sigma > .5$ at level $\alpha = .05$. *Hint: what is the distribution of* $\sum_{i=1}^n (X_i - \bar{X})^2$*, where* $\bar{X} = \frac{1}{n}\sum_{i=1}^n X_i$*?*

3 Problem 3: Maximum likelihood estimation and method of moments

A book has n pages. Each page has a $\text{Pois}(\lambda)$ number of typos, independent of all other pages. Let X_i be the number of typos on page i; we observe $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} \text{Pois}(\lambda)$. Recall that the Poisson PMF is

$$P(X_i = k) = \frac{e^{-\lambda}\lambda^k}{k!}, \quad k = 0, 1, 2, \ldots$$

and the mean and variance are both equal to λ. We are interested in estimating the probability θ that a page has no typos: $\theta = P(X_i = 0) = e^{-\lambda}$.

(a) Consider the MLE for θ, namely, $\hat{\theta}_{\text{MLE}} = e^{-\bar{X}}$, where $\bar{X} = \frac{1}{n}\sum_{i=1}^n X_i$. As $n \to \infty$, $\sqrt{n}(\hat{\theta}_{\text{MLE}} - \theta) \to \mathcal{N}(0, \sigma^2)$. Find σ^2 in terms of λ. *Hint: you can use the delta method.*

(b) Consider a different (perhaps more intuitive) estimator: we simply estimate θ by the fraction of pages with no typos, $\hat{\theta}_{\text{NP}} = \frac{1}{n}(I(X_1 = 0) + I(X_2 = 0) + \cdots + I(X_n = 0))$. We call this estimator $\hat{\theta}_{\text{NP}}$ because it is an example of a *non-parametric* estimator, a type of estimator that makes fewer assumptions about the distribution of the data.

Using the CLT, prove that $\sqrt{n}(\hat{\theta}_{\text{NP}} - \theta) \to \mathcal{N}(0, \tau^2)$ as $n \to \infty$, and find τ^2.

(c) Show that $\hat{\theta}_{\text{NP}}$ is unbiased, and that $\hat{\theta}_{\text{MLE}}$ is not unbiased. *Hint: you may want to use Jensen's inequality which states that for a real-valued random variable Z and a strictly convex function $\phi(\cdot)$ we have $\mathbb{E}[\phi(Z)] \geq \phi(\mathbb{E}[Z])$, with equality only if the variance of Z is zero.*

(d) Show that $\hat{\theta}_{\text{MLE}}$ has a lower asymptotic variance than $\hat{\theta}_{\text{NP}}$, that is show that $\sigma^2 \leq \tau^2$. *Hint: You can use the fact that $e^x \geq 1 + x$, with equality only when $x = 0$.*

4 Problem 4: Likelihood ratio test

Suppose we have one observation $X \in \mathbb{R}$ from the following density:

$$\theta f(x) + (1 - \theta) g(x)$$

where f and g are two distinct densities on $\mathbb{R}$ and $\theta \in [0, 1]$. We further assume that $g(x) > 0$ for all $x \in \mathbb{R}$.

(a) We want to test $H_0 : \theta = 0$ vs $H_1 : \theta = \theta_1$ for some fixed $\theta_1 > 0$. Show that the level-α most powerful test rejects for $\frac{f(X)}{g(X)}$ large enough (you do not need to solve for the threshold).

(b) Is it uniformly most powerful for testing $H_0 : \theta = 0$ vs $H_1 : \theta > 0$?

Fix $\mu > 0$ (assumed to be known in the following). Assume that $f(x)$ is the density of the $\mathcal{N}(\mu, 1)$ distribution, while $g(x)$ is the density of the $\mathcal{N}(0, 1)$ distribution.

(c) Simplify the most powerful test derived in a) so that it has significance level $\alpha = .05$ (and give the threshold).

5 Problem 5: Linear Regression with random X

Background: For linear regression, in the lectures we have treated the outcomes Y_i as random and the covariates X_i as fixed. In this problem, we consider a setting where both Y_i and X_i are random.
Suppose we observe i.i.d. pairs (Y_i, X_i), $i = 1, \ldots, n$, where Y_i measures the amount of ice cream purchased by consumers in one area and X_i measures the average price of ice cream in that area. We assume the linear regression model

$$Y_i = X_i \beta_1 + \beta_0 + \epsilon_i$$

where ϵ_i are i.i.d., $\text{Cov}(\epsilon_i, X_i) = 0$, and $\text{Var}(X_i) \in (0, \infty)$, $\text{Var}(\epsilon_i) \in (0, \infty)$.

Most (but not all) of this problem will be devoted to showing that $\hat{\beta}_1 \xrightarrow{\text{p}} \beta_1$ where

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^n (X_i - \bar{X})^2} \text{ and } \hat{\beta}_0 = \bar{Y} - \bar{X}\hat{\beta}_1$$

are the usual OLS estimators.

(a) Suppose in truth that $\beta_1 \approx 0.5$. Taking into consideration that X_i is random, does this imply mathematically that lowering prices will increase the amount of ice cream purchased? Either prove your answer or give an intuitive example illustrating otherwise.

(b) Show that $\frac{\text{Cov}(Y_1, X_1)}{\text{Var}(X_1)} = \beta_1$.

(c) Show that $\hat{\beta}_1 \xrightarrow{\text{p}} \beta_1$. *Hint: use the law of large numbers to analyze the numerator and denominator of $\hat{\beta}_1$ separately. Then use part (b).*

6 Problem 6: Generalized likelihood ratio test

Suppose that $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} \mathbb{P}$ are i.i.d. draws from some distribution $\mathbb{P}$ on the set of positive integers. For each i, let

$$W_i = \begin{cases} 1 & \text{if } X_i = 1, \\ 2 & \text{if } X_i = 2, \\ 3 & \text{if } X_i \geq 3. \end{cases}$$

Suppose that $n = 1{,}000$ and $X_1, \ldots, X_n$ are unobserved i.i.d. draws from $\mathbb{P}$. Instead we only have access to $W_1, W_2, \ldots, W_n$. The below table summarizes the number of times, out of the sample of size $1,000$, that W_i is either 1, 2, or 3.

	$W_i = 1$	$W_i = 2$	$W_i = 3$
Counts	444	254	302

Suppose we wish to test whether the X_i follow a Geometric distribution distribution. In particular, our null hypothesis H_0 is that $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim}$ Geometric(p) for some $p \in [0,1]$ and let the alternative hypothesis H_A be that $X_1, \ldots, X_n$ are i.i.d., but do not follow a Geometric distribution (i.e. $H_0 : \mathbb{P} = \text{Geometric}(p)$ for some p and $H_A : \mathbb{P} \neq \text{Geometric}(p)$ for any p).

(a) Write down the likelihood of our observed quantities $W_1, \ldots, W_n$, under the null hypothesis, as a function of p, $W_1, W_2, \ldots, W_n$. Note that since $X_1, \ldots, X_n$ are unobserved, we cannot evaluate the likelihood of $X_1, \ldots, X_n$, and therefore such a likelihood cannot be used in a subsequent likelihood ratio test.

(b) Use the table and the likelihood from part (a) to compute the maximum likelihood estimator of p, under the assumption that the null hypothesis H_0 is true.

If you get stuck in (a) or (b), you may use the following intermediate result to solve (c): $\hat{p}_{\text{MLE},0} = \frac{349}{778}$ maximizes the likelihood under the null.

(c) Use your previous results and the data table to conduct a Generalized Likelihood Ratio test of H_0 against H_A at level $\alpha = 0.05$. You must report the generalized likelihood ratio test statistic, the rejection threshold for your test at level $\alpha = 0.05$, and whether or not your test rejects the null at level $\alpha = 0.05$.

Hint: Because only the W_i are observed (and the X_i are not observed) and because the W_i can only take on the value 1, 2 or 3, you can reformulate the alternative hypothesis H_A to be that $W_1, \ldots, W_n$ are i.i.d. and follow some multinomial distribution on $\{1, 2, 3\}$. The test does not have to be exact; you may use an approximation. Quantiles of the chi-square distribution can be found at the end of the exam.

Distribution on X	Support	PDF or PMF	$\mathbb{E}[X]$	$\mathrm{Var}(X)$
Bernoulli(p)	$\{0, 1\}$	$p_X(k) = p^k(1-p)^{1-k}$	p	$p(1-p)$
Binomial(n,p)	$\{0, 1, \dots, n\}$	$p_X(k) = \binom{n}{k} p^k(1-p)^{n-k}$	np	$np(1-p)$
Poisson(λ)	$\{0, 1, 2, \dots\}$	$p_X(k) = \frac{\lambda^k e^{-\lambda}}{k!}$	λ	λ
Geometric(p)	$\{1, 2, 3, \dots\}$	$p_X(k) = p(1-p)^{k-1}$	$\frac{1}{p}$	$\frac{1-p}{p^2}$
Normal(μ,σ^2)	$(-\infty, \infty)$	$f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(x-\mu)^2}$	μ	σ^2
Exponential(λ)	$[0, \infty)$	$f_X(x) = \lambda e^{-\lambda x}$	$\frac{1}{\lambda}$	$\frac{1}{\lambda^2}$
Gamma(α,β)	$[0, \infty)$	$f_X(x) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x}$	$\frac{\alpha}{\beta}$	$\frac{\alpha}{\beta^2}$
Beta(α,β)	$[0, 1]$	$f_X(x) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1}$	$\frac{\alpha}{\alpha+\beta}$	$\frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$

Table 1: Common distributions table of reference. The third column gives the probability mass function (in the case of discrete random variables) or the probability density function (in the case of continuous random variables). Note that the probability mass functions and probability density functions are defined to be equal to zero outside of the support.

Quantiles of the t-distribution

degrees of freedom	$t_{0.9}$	$t_{0.95}$	$t_{0.975}$	$t_{0.99}$	$t_{0.995}$
1	3.08	6.31	12.7	31.8	63.7
2	1.89	2.92	4.30	6.96	9.92
3	1.64	2.35	3.18	4.54	5.84
4	1.53	2.13	2.78	3.75	4.60
5	1.48	2.02	2.57	3.36	4.03
6	1.44	1.94	2.45	3.14	3.71
7	1.41	1.89	2.36	3.00	3.50
8	1.40	1.86	2.31	2.90	3.36
9	1.38	1.83	2.26	2.82	3.25
10	1.37	1.81	2.23	2.76	3.17
11	1.36	1.80	2.20	2.72	3.11
12	1.36	1.78	2.18	2.68	3.05
13	1.35	1.77	2.16	2.65	3.01
14	1.35	1.76	2.14	2.62	2.98
15	1.34	1.75	2.13	2.60	2.95
16	1.34	1.75	2.12	2.58	2.92
17	1.33	1.74	2.11	2.57	2.90
18	1.33	1.73	2.10	2.55	2.88
19	1.33	1.73	2.09	2.54	2.86
20	1.33	1.72	2.09	2.53	2.85
21	1.32	1.72	2.08	2.52	2.83
22	1.32	1.72	2.07	2.51	2.82
23	1.32	1.71	2.07	2.50	2.81
24	1.32	1.71	2.06	2.49	2.80
25	1.32	1.71	2.06	2.49	2.79
26	1.31	1.71	2.06	2.48	2.78
27	1.31	1.70	2.05	2.47	2.77
28	1.31	1.70	2.05	2.47	2.76
29	1.31	1.70	2.05	2.46	2.76
∞	1.28	1.64	1.96	2.33	2.58

For example, to calculate the 95^{th} percentile on one degree of freedom, the first line gives the different quantiles. Choose the entry with column heading $t_{0.95}$ to obtain 6.31.

Quantiles of the chi-square distribution

degrees of freedom	$\chi^2_{0.9}$	$\chi^2_{0.95}$	$\chi^2_{0.975}$	$\chi^2_{0.99}$	$\chi^2_{0.995}$
1	2.71	3.84	5.02	6.63	7.88
2	4.61	5.99	7.38	9.21	10.6
3	6.25	7.81	9.35	11.3	12.8
4	7.78	9.49	11.1	13.3	14.9
5	9.24	11.1	12.8	15.1	16.7
6	10.6	12.6	14.4	16.8	18.5
7	12.0	14.1	16.0	18.5	20.3
8	13.4	15.5	17.5	20.1	22.0
9	14.7	16.9	19.0	21.7	23.6
10	16.0	18.3	20.5	23.2	25.2
11	17.3	19.7	21.9	24.7	26.8
12	18.5	21.0	23.3	26.2	28.3
13	19.8	22.4	24.7	27.7	29.8
14	21.1	23.7	26.1	29.1	31.3
15	22.3	25.0	27.5	30.6	32.8
16	23.5	26.3	28.8	32.0	34.3
17	24.8	27.6	30.2	33.4	35.7
18	26.0	28.9	31.5	34.8	37.2
19	27.2	30.1	32.9	36.2	38.6
20	28.4	31.4	34.2	37.6	40.0
21	29.6	32.7	35.5	38.9	41.4
22	30.8	33.9	36.8	40.3	42.8
23	32.0	35.2	38.1	41.6	44.2
24	33.2	36.4	39.4	43.0	45.6
25	34.4	37.7	40.6	44.3	46.9
26	35.6	38.9	41.9	45.6	48.3
27	36.7	40.1	43.2	47.0	49.6
28	37.9	41.3	44.5	48.3	51.0
29	39.1	42.6	45.7	49.6	52.3
30	40.3	43.8	47.0	50.9	53.7

For example, to calculate the 95^{th} percentile of the chi-square distribution with 30 degrees of freedom, the last line gives the different quantiles. Choose the entry with column heading $\chi^2_{0.95}$ to obtain 43.8.