

## TEST 1: REVIEW SHEET

Here is a list of important results and concepts that might be used in the exam.

### 1 Probability

#### 1.1 Basics of probability

The following concepts are used in the study of statistical quantities. They are like your basic calculus rules and you can use them without justification.

- ☐ Formulas to compute expectation and variance from pdf/pmf
- ☐ Formulas to compute covariance
- ☐ Recognize commonly used distributions by their pdf so you don't have to recompute expectation and variance.
- ☐ Linearity of expectation
- ☐ Expectation of product of independent random variables
- ☐ Variance of average of independent random variables
- ☐ Compute Gaussian probabilities using a table (main trick: standardization)
- ☐ Descriptive statistics: definitions

#### 1.2 Multivariate random variables

This part uses a bit of linear algebra.

- ☐ Covariance matrix: definitions
- ☐ Matrix-vector product, Euclidean inner product and norm
- ☐  $\mathbb{E}[A^\top X]$ ,  $\mathbb{V}(A^\top X)$
- ☐ Multivariate Gaussian distribution properties

### 1.3 Convergence of random variables

These are often used in asymptotic statements about a statistical estimator (consistency, asymptotic normality)

- ☐ Convergence in probability vs. convergence in distribution
- ☐ Law of Large Numbers (LLN)
- ☐ (Univariate and multivariate) Central Limit Theorem (CLT)
- ☐ Delta method (univariate and multivariate)
- ☐ Continuous Mapping Theorem (CMT)
- ☐ Slutsky's theorem
- ☐ Operations on convergence results (addition, multiplication, division)

## 2 Statistical inference

In this section, we list the important concepts of statistical inference that you should know (definition and how to compute using above probability rules).

### 2.1 Statistical models

- ☐ Parametric vs. nonparametric

### 2.2 Point estimation

- ☐ Estimator
- ☐ Bias (+ unbiased, asymptotically unbiased)
- ☐ Standard error
- ☐ MSE
- ☐ Consistency
- ☐ Asymptotic normality, asymptotic variance
- ☐ Constructing confidence intervals via CLT

# EXERCISE

Let  $X$  be a multivariate Gaussian random vector with unknown mean  $\mu = (\mu_1, \mu_2) \in \mathbb{R}^2$  and covariance matrix

$$\Sigma = \begin{pmatrix} 4 & 0 \\ 0 & 5 \end{pmatrix}.$$

Define  $Y = X^{(1)} + X^{(2)}$  and  $Z = X^{(1)}X^{(2)}$ , where  $X = (X^{(1)}, X^{(2)})$ . We would like to estimate the parameter  $\theta = \|\mu\|^2$ . To that end, we observe  $X_1, \dots, X_n$  which are i.i.d. with the same distribution as  $X$  and we propose the estimator

$$\hat{\theta} = \bar{Y}_n^2 - 2\bar{Z}_n,$$

where

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

and

$$\bar{Z}_n = \frac{1}{n} \sum_{i=1}^n X_i^{(1)} X_i^{(2)}$$

1. Compute the mean and covariance matrix of  $(Y, Z)$ .

**Solution.** We have that

$$\mathbb{E}[Y] = \mathbb{E}[X^{(1)} + X^{(2)}] = \mu_1 + \mu_2$$

and

$$\mathbb{E}[Z] = \text{Cov}(X^{(1)}, X^{(2)}) + \mathbb{E}[X^{(1)}]\mathbb{E}[X^{(2)}] = \mu_1\mu_2.$$

For the covariance matrix, we must compute  $\mathbb{V}(Y)$ ,  $\mathbb{V}(Z)$ , and  $\text{Cov}(Y, Z)$ . We have

$$\mathbb{V}(Y) = \mathbb{V}(X^{(1)}) + \mathbb{V}(X^{(2)}) + 2\text{Cov}(X^{(1)}, X^{(2)}) = 4 + 5 + 0 = 9$$

and

$$\begin{aligned} \mathbb{V}(Z) &= \mathbb{E}[(X^{(1)}X^{(2)})^2] - \mathbb{E}[X^{(1)}X^{(2)}]^2 \\ &= \mathbb{E}[(X^{(1)})^2]\mathbb{E}[(X^{(2)})^2] - (\mu_1\mu_2)^2 \\ &= (4 + \mu_1^2)(5 + \mu_2^2) - (\mu_1\mu_2)^2 \\ &= 20 + 5\mu_1^2 + 4\mu_2^2 \end{aligned}$$

where we used independence to simplify the calculation (recall that zero correlation in normal random variables implies independence). If we had correlation (i.e. non-zero entries on the off-diagonal entries of  $\Sigma$ ), we could do this calculation by writing  $X^{(1)}$  and  $X^{(2)}$  as linear combinations of  $Z_1$  and  $Z_2$  where  $Z_1$  and  $Z_2$  are standard normals, such that the

values of  $\mu$  and  $\Sigma$  match.

Finally, we have

$$\begin{aligned}\text{Cov}(Y, Z) &= \mathbb{E}[(X^{(1)} + X^{(2)})X^{(1)}X^{(2)}] - (\mu_1 + \mu_2)\mu_1\mu_2 \\ &= (4 + \mu_1^2)\mu_2 + (5 + \mu_2^2)\mu_1 - (\mu_1 + \mu_2)\mu_1\mu_2 \\ &= 4\mu_2 + 5\mu_1\end{aligned}$$

Putting this all together, we have

$$\mathbb{E}[(Y, Z)] = (\mu_1 + \mu_2, \mu_1\mu_2)$$

and

$$\mathbb{V}[(Y, Z)] = \begin{pmatrix} 9 & 5\mu_1 + 4\mu_2 \\ 5\mu_1 + 4\mu_2 & 20 + 5\mu_1^2 + 4\mu_2^2 \end{pmatrix}.$$

**Note:** if  $Z$  was instead a linear combination of  $X^{(1)}$  and  $X^{(2)}$ , we can write the vector  $(Y, Z)$  as  $AX$ , where  $A$  is a  $2 \times 2$  matrix. We can calculate the mean and covariance matrix using  $\mathbb{E}[AX] = A\mathbb{E}[X]$  and  $\mathbb{V}[AX] = A\Sigma A^\top$ .

2. Is  $\hat{\theta}$  consistent? Why or why not?

**Solution.** Yes. By LLN, note that  $\bar{Y}_n$  is consistent for  $\mu_1 + \mu_2$  and  $\bar{Z}_n$  is consistent for  $\mu_1\mu_2$ . Therefore, by **Continuous Mapping Theorem** and the **addition rule for convergence in probability**, we have that  $\bar{Y}_n^2 - 2\bar{Z}_n$  is consistent for  $(\mu_1 + \mu_2)^2 - 2\mu_1\mu_2 = \mu_1^2 + \mu_2^2$ .

3. Compute the bias of  $\hat{\theta}$ . Is it biased, unbiased, or asymptotically unbiased?

**Solution. Method 1.**

$$\begin{aligned}\mathbb{E}[\bar{Y}_n^2 - 2\bar{Z}_n] &= \mathbb{V}(\bar{Y}_n) + \mathbb{E}[\bar{Y}_n]^2 - 2\mathbb{E}[\bar{Z}_n] \\ &= \frac{1}{n}\mathbb{V}(Y_1) + \mathbb{E}[Y_1]^2 - 2\mathbb{E}[Z_1] \\ &= \frac{9}{n} + (\mu_1 + \mu_2)^2 - 2\mu_1\mu_2 \\ &= \frac{9}{n} + \mu_1^2 + \mu_2^2\end{aligned}$$

**Method 2.** We have that

$$\begin{aligned}
\mathbb{E}[\bar{Y}_n^2 - 2\bar{Z}_n] &= \frac{1}{n^2} \mathbb{E} \left[ \left( \sum_{i=1}^n X_i^{(1)} + X_i^{(2)} \right)^2 \right] - \frac{2}{n} \mathbb{E} \left[ \sum_{i=1}^n X_i^{(1)} X_i^{(2)} \right] \\
&= \frac{1}{n^2} \mathbb{E} \left[ \sum_{i=1}^n (X_i^{(1)} + X_i^{(2)})^2 + \sum_{i \neq j} (X_i^{(1)} + X_i^{(2)})(X_j^{(1)} + X_j^{(2)}) \right] - \frac{2}{n} \cdot n\mu_1\mu_2 \\
&= \frac{1}{n^2} \mathbb{E} \left[ n(X_1^{(1)} + X_1^{(2)})^2 + n(n-1)(X_1^{(1)} + X_1^{(2)})(X_2^{(1)} + X_2^{(2)}) \right] - 2\mu_1\mu_2 \\
&= \frac{1}{n^2} \mathbb{E} \left[ n(4 + \mu_1^2 + 2\mu_1\mu_2 + 5 + \mu_2^2) + n(n-1)(\mu_1 + \mu_2)^2 \right] - 2\mu_1\mu_2 \\
&= \frac{9}{n} + \mu_1^2 + \mu_2^2.
\end{aligned}$$

With either method,  $\mathbb{E}[\hat{\theta}] - \theta = \frac{9}{n}$  and the estimator is biased. Since the bias goes to zero when  $n \rightarrow \infty$ , it will also be asymptotically unbiased.

4. Write a multivariate central limit theorem for  $(\bar{Y}_n, \bar{Z}_n)$ .

**Solution.** Using the mean and covariance we calculated in part 1, we have

$$\sqrt{n} \left( \begin{pmatrix} \bar{Y}_n \\ \bar{Z}_n \end{pmatrix} - \begin{pmatrix} \mu_1 + \mu_2 \\ \mu_1\mu_2 \end{pmatrix} \right) \rightsquigarrow \mathcal{N} \left( 0, \begin{pmatrix} 9 & 5\mu_1 + 4\mu_2 \\ 5\mu_1 + 4\mu_2 & 20 + 5\mu_1^2 + 4\mu_2^2 \end{pmatrix} \right)$$

5. Show that  $\hat{\theta}$  is asymptotically normal and compute its asymptotic variance  $\sigma^2$ .

**Solution.** We can use multivariate Delta method, with  $g(y, z) = y^2 - 2z$ . Note that  $\hat{\theta} = g(\bar{Y}_n, \bar{Z}_n)$  for this choice of  $g$ . We have that  $g$  is differentiable and  $\nabla g(y, z) = (2y, -2)^\top$ . Now, we may apply the Delta method:

$$\begin{aligned}
\sqrt{n}(g(\bar{Y}_n, \bar{Z}_n) - g(\mu_1 + \mu_2, \mu_1\mu_2)) &\rightsquigarrow \\
&\mathcal{N} \left( 0, (2\mu_1 + 2\mu_2, -2) \begin{pmatrix} 9 & 5\mu_1 + 4\mu_2 \\ 5\mu_1 + 4\mu_2 & 20 + 5\mu_1^2 + 4\mu_2^2 \end{pmatrix} \begin{pmatrix} 2\mu_1 + 2\mu_2 \\ -2 \end{pmatrix} \right).
\end{aligned}$$

Simplifying gives

$$\sqrt{n}(\hat{\theta} - \theta) \rightsquigarrow \mathcal{N}(0, 16\mu_1^2 + 20\mu_2^2 + 80).$$

Therefore,  $\hat{\theta}$  is asymptotically normal with asymptotic variance  $16\mu_1^2 + 20\mu_2^2 + 80$ .

6. Propose a consistent estimator  $\hat{\sigma}^2$  of  $\sigma^2$ , and write a 90% confidence interval for  $\theta$  that is symmetric about  $\hat{\theta}$ .

**Solution.** A consistent estimator for  $\mu_1$  is  $\bar{X}_n^{(1)} = \frac{1}{n} \sum_{i=1}^n X_i^{(1)}$ , and a consistent estimator for  $\mu_2$  is  $\bar{X}_n^{(2)} = \frac{1}{n} \sum_{i=1}^n X_i^{(2)}$ . Therefore, by continuous mapping theorem, a consistent

estimator for  $\sigma^2$  is  $\hat{\sigma}^2 = 16 \left( \bar{X}_n^{(1)} \right)^2 + 20 \left( \bar{X}_n^{(2)} \right)^2 + 80$ .

We would like a  $1 - \alpha = 90\%$  symmetric confidence interval. We have  $\Phi(1.65) \approx .95 = 1 - \frac{\alpha}{2}$ , so a symmetric 90% confidence interval is

$$\left( \hat{\theta} - 1.65 \frac{\sqrt{16 \left( \bar{X}_n^{(1)} \right)^2 + 20 \left( \bar{X}_n^{(2)} \right)^2 + 80}}{\sqrt{n}}, \hat{\theta} + 1.65 \frac{\sqrt{16 \left( \bar{X}_n^{(1)} \right)^2 + 20 \left( \bar{X}_n^{(2)} \right)^2 + 80}}{\sqrt{n}} \right).$$

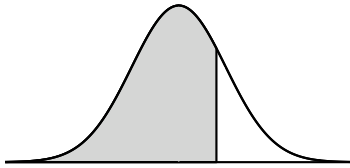


Table 1: The table lists  $P(Z \leq z)$  where  $Z \sim N(0, 1)$  for positive values of  $z$ .

$Z$	Second decimal place of $Z$									
	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
3.0	0.9987	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.9990	0.9990
3.1	0.9990	0.9991	0.9991	0.9991	0.9992	0.9992	0.9992	0.9992	0.9993	0.9993
3.2	0.9993	0.9993	0.9994	0.9994	0.9994	0.9994	0.9994	0.9995	0.9995	0.9995
3.3	0.9995	0.9995	0.9995	0.9996	0.9996	0.9996	0.9996	0.9996	0.9996	0.9997
3.4	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9998

\*For  $Z \geq 3.50$ , the probability is greater than or equal to 0.9998.