SMDS - Collection of exercises to be solved in the class

Convergence and limit theorems

Exercise 1. A teacher has to grade 50 exams in sequence. The times required to grade the 50 exams are independent with a common distribution that has mean 20 minutes and standard deviation 4 minutes. Find an approximation of the probability that the teacher will grade at least 25 of the exams within 450 minutes.

Exercise 2. A balanced dice is rolled 100 times. Find the probability that the sum of results is between 320 and 380 (extremes included).

Exercise 3. Law of Large Numbers states that, given a sequence X_i of random variables independent and identically distributed, the resulting sequence of sample means $\overline{X}_n$ converges in probability to their mean $E[X_i] = \mu$. Determine the limit of the sequence $\{\overline{Y}_n\}$ of geometrical means, where

$$\overline{Y}_n = \Big(\prod_{i=1}^{n} X_i\Big)^{1/n}.$$

Exercise 4. A multinational tobacco corporation claims that the amount of nicotine in one of its cigarettes has a mean 2.2mg and standard deviation 0.3mg. Randomly choosing 100 cigarettes the observed mean is 3.1mg. Approximate the probability for the sample mean to be equal or greater than the one observed, if the real mean is 2.2mg.

Exercise 5. Let X be a variable with both mean and variance equal to 20. Using a notable inequality find a lower bound for $P[0 < X < 40]$.

Exercise 6. A man has 100 light bulbs whose lifetimes are exponentially distributed with mean 5 hours. If these are used one by one in sequence (replaced when the used one gets out of order), approximate the probability that there will be a light bulb still working after 525 hours.

Exercise 7. An insurance company has 10000 clients. Refund requests from every single client have mean 240 dollars and standard deviation 800 dollars.

Approximate the probabilitiy that total refund request in a year will be greater than 2.7 million dollars.

Exercise 8. Let $X_1, X_2, \ldots, X_{30}$ be random variables with Poisson distribution of mean 1. Let $S = \sum_{i=1}^{30} X_i$.

a) Using Markov inequality find a bound for the probability $P[S > 25]$;

b) Using CLT (Central Limit Theorem) approximate $P[S > 25]$.

Exercise 9. Let X be a discrete random variable with values in $\mathbf{N}^+ = \{1, 2, \ldots\}$. Show that, if $P[X = k]$ is decresing (not strictly) in k, then

$$P[X = k] \leq 2\frac{E[X]}{k^2}$$

Point estimations

Exercise 10. Let $y_1 = 0.42, y_2 = 0.10, y_3 = 0.65$ and $y_4 = 0.23$ be a sample with dimension 4 of a random variable Y with density

$$f_Y(y, \theta) = \theta y^{\theta - 1}, \qquad y \in [0, 1], \theta > 0.$$

Estimate θ through method of moments.

Exercise 11. Let $y_1 = 3, y_2 = 5, y_3 = 4$ and $y_4 = 2$ a random sample of dimension 4 of the variable Y with Poisson distribution, i.e.

$$f_Y(k, \lambda) = e^{-\lambda}\frac{\lambda^k}{k!}, \qquad k \in \mathbf{N}.$$

Estimate λ with maximum likelihood method.

Exercise 12. Suppose an isolated weather-reporting station has an electronic device whose time to failure is given by the exponential model

$$f_Y(y, \theta) = \frac{1}{\theta} e^{-1/\theta}, \qquad y \in \mathbf{R}^+.$$

The station also has a spare device, so the time until this instrument is not available is the sum S of these two exponentials variables, which is

$$f_S(y, \theta) = \frac{1}{\theta^2} y e^{-1/\theta}, \qquad y \in \mathbf{R}^+.$$

Five data points of S have been collected, with values 9.2, 5.6, 18.4, 12.1 and 10.7. Find the maximum likelihood estimate for θ.

Exercise 13. Let $X_1, X_2, \ldots, X_n$ a sample, recording the number of yearly claims (for n consecutive years) that a condominium have issued to an insurance company, for some adverse event that could be covered by its policy. Suppose that this data-set can be modeled as a simple random sample from a random variable X with Poisson distribution.

a) Find estimators of λ through method of moment and method of maximum likelihood.

b) In order to find the estimator of λ through Bayesian inference, adopt a Gamma prior and prove that this family of distributions is conjugate to the Poisson family. Suppose that the company experience is that most of the policyholders issue one or two claims per year (rarely 0 or 3, other cases hava negligible probabilities). Try to encode that information into the choice of the prior and exhibit a Bayesian point estimator that takes this information into account. What can you say about the distribution fo this estimators?

Exercise 14. Let $X_1, X_2, \ldots, X_n$ be a sample of lifetimes of a device. Suppose that X has an exponential distribution with parameter λ.

a) Find estimators of λ through method of moment and method of maximum likelihood.

b) In order to perform Bayesian inference, adopt a Gamma prior and prove that this family of distributions is conjugate to the Exponential family.

Exercise 15. Let X a variable with pdf

$$f_X(x, \theta) = \theta x^{\theta-1}, \qquad x \in [0, 1], \theta > 0.$$

Estimate θ through method of maximum likelihood. medskip

Exercise 16. Each item produced by a company will, independently, be defective with probability p. If the prior distribution on p is uniform on (0,1), compute the posterior probability that p is less than 0.2 given

a) a total of 2 defective out of a sample of size 10;

b) a total of 1 defective out of a sample of size 10;

c) a total of 10 defective out of a sample of size 10;

Exercise 17. Using method of moment estimate θ if pdf is

$$f_X(x,\theta) = (\theta^2 + \theta)x^{\theta-1}(1-x) \qquad x \in [0,1]$$

Assume you have a sample of size n.

Exercise 18. Let $y_1 = 8.3, y_2 = 4.9, y_3 = 2.6$ and $y_4 = 6.5$ a sample of dimension 4 from a uniform distribution depending on two parameters

$$f_Y(y,\theta_1,\theta_2) = \frac{1}{2\theta_2} \qquad \theta_1 - \theta_2 \le y \le \theta_1 + \theta_2.$$

Estimate both parameters through method of moment.

Exercise 19. Find a formula for the method of moments estimate for the parameter θ in the Pareto pdf

$$f_Y(y,\theta) = \theta k^\theta \big(1/y\big)^{\theta+1}, \quad y \ge k, \quad \theta \ge 1.$$

Assume that k is known and that data consist of a random sample of size n.

Exercise 20. Calculate the method of moments estimate for the parameter θ in the pdf

$$p_X(k,\theta) = \theta^k(1-\theta)^{1-k}, \qquad k = 0,1,$$

if the sample size is $n = 5$ and observed values are $0, 0, 1, 0, 1$.

Exercise 21. Calculate with the maximum likelihood method the estimate of the parameter θ of the discrete density

$$p_X(k,\theta) = \theta^k(1-\theta)^{1-k}, \qquad k = 0,1,$$

when $n = 8$ and the observed values are $0, 0, 1, 0, 1, 1, 1, 1$.

Exercise 22. Use the sample $y_1 = 8.2, y_2 = 9.1, y_3 = 10.6$ and $y_4 = 4.9$ to compute maximum likelihood estimate for λ in exponential pdf

$$f_Y(y,\lambda) = \lambda e^{-\lambda y}, \qquad y \ge 0.$$

Exercise 23. Suppose a random sample of size n is drawn from the probability model

$$p_X(k,\theta) = \frac{\theta^{2k}e^{-\theta^2}}{k!}, \qquad k = 0, 1, 2, ...$$

Find a formula for the maximum likelihood estimator for θ.

Exercise 24. Calculate the maximum likelihood estimator for the parameter θ of the pdf

$$f_X(x,\theta) = \frac{x^3 e^{-x/\theta}}{6\theta^4}, \quad y \geq 0,$$

given $n = 3$ observed values $2.3, 1.9, 4.6$.

Exercise 25. Calculate an estimate for the parameter θ of pdf

$$f_X(x,\theta) = \theta x^{\theta-1}, \quad 0 \leq x \leq 1, \quad \theta > 0,$$

given $n = 5$ observed values $0.77, 0.82, 0.92, 0.94, 0.98$.

Exercise 26. Find the maximum likelihood estimate for θ in the pdf

$$f_Y(y,\theta) = \frac{2y}{1-\theta^2}, \quad \theta \leq y \leq 1,$$

if a random sample of size $n = 6$ yielded the measurements $0.70, 0.63, 0.92, 0.86, 0.43, 0.21$.

Exercise 27. Suppose taht X is a geometric random variable with $p_X(k,\theta) = (1-\theta)^{k-1}\theta,\ k = 1, 2, ...$ Assume that the prior distribution for θ is the Beta pdf with parameters r and s. Find the posterior distribution for θ given a sample of size $n = 1$.

Exercise 28. Suppose a coin, for which probability p of getting head is unknown, is to be tossed ten times for the purpose of estimating p with the function $\hat{p} = X/10$ where X is the observed number of heads.

a) If $p = 0.60$, what is the probability that $|\hat{p} - 0.60| \leq 0.10$? That is, what are the chances that the estimator will fail within 0.10 of the true value of the parameter?

b) Repeat point (a) assuming coin is tossed 100 times.

Exercise 29. Let $\hat{\theta}_1 = \frac{3}{2}\overline{X}$ and $\hat{\theta}_2 = X_{\max}$ be two estimators of θ in the pdf $f_X(x,\theta) = \frac{2x}{\theta^2}$, $0 \le x \le \theta$. Are either or both unbiased?

Exercise 30. Let $X_1, X_2, ..., X_n$ be as random sample from the pdf $f_X(x,\theta) = \frac{2x}{\theta^2}$, $0 \le x \le \theta$. We know that $\hat{\theta}_1 = \frac{3}{2}\overline{X}$ and $\hat{\theta}_2 = \frac{2n+1}{2n}X_{\max}$ two unbiased estimators for θ. Which estimator is more efficient?

Exercise 31. Let X be a random variable with Binomial distribution of parameters n and p, and let $\hat{p} = X/n$ be an unbiased estimator of p. Compare $V[\hat{p}]$ with the Cramer-Rao lower bound for unbiased estimator of p.

Exercise 32. Consider uniform pdf $f_Y(y,\theta) = 1/\theta$, $0 \le y \le \theta$, let $\hat{\theta}_n = Y_{\max}$ an estimator of θ, which we know is unbiased. Say if it's also consistent.

Exercise 33. A sample of size $n = 16$ is drawn from a normal distribution with standard deviation $\sigma = 10$ and mean μ unknown. If $\mu = 20$, what is the probability that the estimator $\hat{\mu} = \overline{X}_n$ will lie between 19.0 and 21.0?

Exercise 34. A random sample is made up of Y_1 and Y_2, drawn from a pdf

$$f_Y(y,\theta) = 2y\theta^2, \quad 0 < y < 1/\theta.$$

What must c equal if the statistics $c(Y_1 + 2Y_2)$ is to be an unbiased estimator of $1/\theta$.

Exercise 35. Let $X_1, X_2, ..., X_n$ be a random sample from the pdf $f_X(x,\theta) = \frac{1}{\theta}e^{-x/\theta}$, $x > 0$.

a) Show that $\hat{\theta}_1 = X_1$, $\hat{\theta}_2 = \overline{X}_n$ and $\hat{\theta}_3 = n \cdot X_{\min}$ are unbiased estimators for θ.

b) Find variances of the three estimators.

c) Calculate the relative efficiency of the three estimators.

Exercise 36. Let $X_1, X_2, ..., X_n$ be a random sample from pdf $f_X(x,\theta) = \frac{1}{\theta}e^{-x/\theta}$, $x > 0$. Compare the Cramer-Rao lower bound for unbiased estimators of θ to the variance of the maximum likelihood estimator for θ, $\hat{\theta} = \overline{X}_n$. Is $\overline{X}_n$ a best estimator for θ?

Exercise 37. Let $X_1, X_2, ..., X_n$ be a sample drawn from a normal population with mean μ and variance σ^2, where σ^2 is known. Compare the Cramer-Rao lower bound for unbiased estimator of μ with the variance of $\hat{\mu} = \overline{X}_n$. Is $\hat{\mu}$ an efficient estimator?

Exercise 38. How large a sample must be taken from a normal pdf with mean $\mu = 18$ and variance $\sigma^2 = 25$ in order to guarantee that $\hat{\mu} = \overline{X}_n$ has a 90% probability of lying in the interval [16, 20]?

Interval estimation

Exercise 39. The principal randomly selected six students out of 100 to take an aptitude test (whose possible score is in $[0, 100] \subseteq \mathbb{R}$). Their scores where: 89.8, 76.3, 79.4, 87.6, 79.4, 70.9. Assuming scores have normal distribution, determine a 90% confidence interval for the mean score for all 100 students.

Exercise 40. A population is normal with variance 68. Suppose you want to estimate the mean μ. Find the sample size needed to assure with 68% confidence that the sample mean will not differ from the population mean by more than 3.

Exercise 41. A population is normal with standard deviation 100. Suppose you want to estimate the mean μ. If a sample of dimension 25 gives a sample mean 8439, determinate correspondent 90% and 92% confidence intervals.

Exercise 42. Given the sample $\{5.2, 6.4, 4.8, 5.0, 5.6, 4.8\}$ drawn from a normal distribution, provide an interval estimate:

a) of the mean with confidence level $\alpha = 0.05$;

b) of the variance with a confidence level $\alpha = 0.1$.

Quiz 43. Given a sample of 4 measures extracted from a population with normal distribution, we obtain a sample mean equal to 10 and a sample corrected variance equal to 8. Then the 95% confidence interval for the mean is given by:

a) $(10 - t_{4,0.975} \cdot \sqrt{2}, 10 + t_{4,0.975} \cdot \sqrt{2})$

b) $(10 - t_{3,0.975} \cdot \sqrt{2}, 10 + t_{3,0.975} \cdot \sqrt{2})$

c) $(10 - z_{0.975} \cdot \sqrt{2}, 10 + z_{0.975} \cdot \sqrt{2})$

Quiz 44. Given the sample $\{12, 14, 16\}$, then the confidence interval for the variance

a) can be always provided

b) can be provided only if the population is normally distributed

c) can be never provided

Quiz 45. Let $X_1, X_2, ..., X_n$ be a random sample drawn from a normal pdf with mean μ and variance σ^2, both unknown. The amplitude of the interval with confidence level α for the mean

a) does not dipend on sample variance;

b) can not be determined due to some missing information;

c) depends on sample mean's value;

d) is slightly larger than the one that could be obtained (improperly) assuming the variance is known and equal to the value estimated from the sample;

e) does not dipend on the sample size n.

Quiz 46. Let $X_1, X_2, ..., X_n$ and $Y_1, Y_2, ..., Y_n$ be two independent samples from two normal population both with variance σ^2 known. The amplitude of the confidence interval on the difference between the two means

a) can not be determined due to some missing information;

b) does not depend on sample size n;

c) reduces as $1/\sqrt{n}$ as the sample size n increases;

d) grows with sample size n;

e) depends on sample means $\overline{X}_n$ and $\overline{Y}_n$.

Quiz 47. Let $X_1, X_2, ..., X_8$ and $Y_1, Y_2, ..., Y_7$ be two samples from two normal distributions with the same unknown variance σ^2. Let $\overline{X}_8 = 123.1, \hat{\sigma}_X = 8.4, \overline{Y}_7 = 119.3$ and $\hat{\sigma}_Y = 11.7$. A confidence interval for the difference $\mu_Y - \mu_X$ is

a) [2.23, 5.82];

b) [-7.45, 15.04];

c) a couple of numbers such that the probability that the parameter lies between them is fixed;

d) a couple of numbers whose difference has the role of an estimate of the difference between the two means;

e) [-18.40, 23,56].

Exercise 48. Suppose you have a sample $X_1, X_2, ..., X_n$ drawn from a normal distribution of mean μ and variance σ^2. Denote X_{n+1} as a further observation (that is not available now. A prediction interval of level α is an interval in which you would expect X_{n+1} to lie with probability $1 - \alpha$. It can be found in the same way you determine a confidence interval.

a) What can you say on the distribution of $X_{n+1} - \overline{X}$?

b) Suppose first that the variance is known, and then try to generalize to the case when it is unknown. Can you use this information to find a prediction interval for X_{n+1}?

Exercise 49. It is important that face masks used by fire fighters be able to withstand high temperatures because the commonly work in temperatures of 200°-500° F. In a test of one type of mash, 11 out of 55 had lenses pop out at 250°. Construct a 90% confidence interval for the true proportion of masks of this type whose lenses would pop out at 250°.

Exercise 50. A random sample of $n = 15$ heat pumps of a certain type yielded the following observations in lifetime (in years):
2.0, 1.3, 6.0, 1.9, 5.1, 0.4, 1.0, 5.3, 15.7, 0.7, 4.8, 0.9, 12.2, 5.3, 0.6.

a) Assume that the lifetime distribution is exponential and obtain a 95% confidence interval for expected (true average) lifetime.

b) Repeat to find a 98% confidence interval.

Exercise 51. A sample of dimension 9 observed from a normal distribution assumed the following values:
0.2, 1.4, 2.3, 0.6, 2.5, -1.3, 0.8, -1.8, -0.2.

a) Provide an interval estimate of the mean with a confidence level $\alpha = 0.1$, assuming the variance is known and equal to 4;

b) Provide an interval estimate of the mean with confidence level $\alpha = 0.1$, assuming variance is unknown;

c) Provide an interval estimate of the variance with confidence level $\alpha = 0.1$.

Exercise 52. Suppose a random sample of size $n = 11$ is drawn from a population Y normally distributed with mean $\mu = 15.0$. For what value of k it's true that

$$P\left(\left|\frac{\overline{Y} - 15.0}{S/\sqrt{11}}\right| \geq k\right) = 0.05$$

Hypothesis testing

Exercise 53. To evaluate the carbon dioxide emitted from the cars we produce, a number of 50 different measures (from 50 different vehicles) are performed and recorded. We obtain a sample mean equal to 12.5 kt and a sample variance equal to 0.36.

a) Provide a confidence interval form the mean with confidence 99%;

b) Test, with confidence 95%, the hypothesis that the true mean is 12.0 kt.

Exercise 54. To evaluate the carbon dioxide emitted from the cars we produce, a number of 5 different measures (from 5 different vehicles) are performed and recorded. They are listed below:

$$12.5, 11.1, 13.4, 11.9, 13.1$$

(in kt). Assume that the emissions are normally distributed.

a) Provide a confidence interval for the mean with confidence 99%;

b) Test, with confidence 95%, the hypothesis that the true mean is 12.0 kt.

Exercise 55. To evaluate the carbon dioxide emitted from the cars we produce, a number of 5 different measures (from 5 different vehicles) are performed and recorded. They are listed below:

$$12.5, 11.1, 13.4, 11.9, 13.1$$

(in kt). Assume that the emissions are normally distributed.

a) Provide a confidence interval for the variance with confidence 95%;

b) Test, with confidence 95%, the hypothesis that the true variance is 0.5.

Exercise 56. Given the sample $\{0.15, 1.25, -0.55, 0.85, 2.00, -1.30\}$ drawn from a normal distribution,

a) test the hypothesis that the variance is equal to 2,00 with level $\alpha = 0.05$;

b) To test the hypothesis $H_0 : \mu = 0$ the acceptance region $I = (-1, 1)$ is fixed (and the sample mean as the test statistic). Given the dimension $n = 10$ for the sample, and assuming variance known and equal to 2, what is the probability of an error of the I kind?

Exercise 57. Natural cork in wine bottles is subject to deterioration and as a result wine in such bottles may experience contamination. The article "Effects of Bottle Closure Type on Consumer Perceptions of Wine Quality" (Amer. J. of Enology and Viticulture, 2007: 182–191) reported that, in a tasting of commercial Chardonnays, 16 of 91 bottles were considered spoiled to some extent by cork-associated characteristics. Does this data provide strong evidence that more than 15% of all such bottles are contaminated in this way? Let's carry out a test of hypothesis using a significance level of 0.10.

Exercise 58. Let the price paid for a given common good by different public administration be approximately normal with a mean 34.30. Suppose the same good has been bought 26 times by Politecnico di Torino in the past year, at prices with an average 39.54 and with a standard deviation of 5.2 euros. We may suspect that the price paid by our University is anomalously high and hides some misbehavior by the administration. Is there enough evidence (with confidence level $\alpha = 0.05$) to exclude that such an higher mean is only due to random fluctuations? Could we keep the same decision it the requested confidence level was $\alpha = 0.01$? How do we compute the P-value?

Exercise 59. The prevalence of a disease in a population is the probability that a randomly chosen individual in the population is affected by the disease. Let us focus on a specific disease whose prevalence is $p = 2.7 \cdot 10^{-4}$ in the general population. Suppose that we want to check whether in a specific region a source of pollutants may raise the prevalence of the disease. We pick all 42500 inhabitants of the region and record 14 people affected. Is it enough to conclude that in the region the prevalence of the disease is higher than in the general population?

Exercise 60. A sample of 50 lenses used in eyeglasses yields a sample mean of thickness of 3.05 mm and a sample standard deviation of 0.34 mm. The desidered true average thickness of such lenses is 3.20 mm. Does the data strongly suggest that the true average thickness of such lenses is something other than what is desired? Test using $\alpha = 0.05$.

Exercise 61. A random sample of 150 recent donations at a certain blood bank reveals that 82 were type A blood. Does this suggest that the actual percentage of type A donations differs from 40%, the percentage of the population having type A blood? Test the hypothesis using a significance level of 0.01. Would your conclusion have been different if a significance level 0.05 had been used?

Exercise 62. The manufacturer of a particular type of aluminium bars states that they present a number of defects, for each meter, that has Poisson distribution with mean $\lambda = 1$. To test his assertion, we check 20 pieces whose length is 1 meter, obtaining the following: 0 flaws in 5 cases, 1 flaw in 7 cases, 2 flaws in 3 cases, more than 2 flaws in the remaining 5 cases. Test, with the χ^2 method and confidence level 95%, if what he affirms can not be rejected.

Exercise 63. The following data relate the mother's age and the birth weight (in grams) of her child: for mothers with 20 years or less we observe 10 cases in which birth weight is less than 2500 g whereas in 40 cases the birth weight is more than 2500 g. For mothers with more than 20 years we observe 15 cases in which baby's weight is less then 2500 g whereas in 135 cases the weight is more than 2500 g. Test the hypothesis that the baby's birth weight is independent of the mother's age with confidence level $\alpha = 0.01$.

Exercise 64. We want to test with level of significance 0.05 the hypothesis that a given population is uniformly distributed on $[0, 10]$. What is the conclusion if the ordered values from a sample size 10 are the following?

$$1.2, 1.5, 2.3, 2.4, 4.0, 5.5, 6.1, 7.8, 8.1, 9.5.$$

Exercise 65. A sample of size 120 had a sample mean of 100 and a sample standard deviation of 15. Observed values are summarized in this table

<70	70-85	85-100	100-115	115-130	>130
3	18	30	35	32	2

Test the hypothesis that the sample distribution is normal.

Exercise 66. An experiment is designed to study the possible relationship between hypertension and cigarette smoking. A sample made up of 190 people randomly choosen yielded the following data

	Nonsmoker	**Moderate smoker**	**Heavy smoker**
Hypertension	20	38	28
No hypertension	50	27	18

Test the hypothesis that whether or not an individual has hypertension is independent of how much that person smokes with level of significance 0.10.

Exercise 67. A sample of 120 size has a sample mean equal to 100 and a sample adjusted variance equal to 15. Of the 120 observed values, 3 have a value less than 70, 18 between 70 and 85, 30 between 85 and 100, 35 between 100 and 115, 32 between 115 and 130, and 2 have value greater than 130. Test the hypothesis that the distribution is normal with 95% significance.

Exercise 68. We are conducting a study to understand if there is a relationship between hypertension and smoking. The data for 190 randomly extracted individuals are then analyzed. Of them, 85 have hypertension while 95 have not. Among those suffering from hypertension, 20 are non-smokers, 38 moderate smokers and 27 heavy smokers. Among those who do not suffer from hypertension, 50 are non-smokers, 27 moderate smokers and 18 heavy smokers. Test the independence between hypertension and smoking (on three levels) with a level of significance equal to 0.10.

Recap exercises

Quiz 71. After a large clinical trial (1500 patients) it has been estimated (in the sense of point estimation) that the probability of developing a specific side effetc is 0.0173. Using normal approximation, which one is correct?

a) an approximate unilateral 95% confidence interval for p is $(-\infty, 0, 0482)$;

b) an approximate unilateral 95% confidence interval for p is $(-\infty, 0, 0228)$;

c) an approximate bilateral 95% confidence interval for p is $(0, 0023, 0, 0273)$;

d) none of the other option is correct;

e) an approximate bilateral 95% confidence interval for p is p è $(0, 0168, 0, 0178)$.

Quiz 72. If you want to reduce the expected amplitude of the confidence interval for the mean of a normal sample by a factor of 2 ...

a) ... you need to collect a sample with sample size 4 times higher than the current one

b) ... none of the other options is correct

c) ... you need to collect a sample with double sample size

d) ... you need to collect a sample with sample size 4 times smaller then the current one

e) ... you need to reduce the sample size by a factor 2

Quiz 73. For security reasons the body temperature of all people entering Politecnico has been measured for the past two months, recording a total of 3634 measurements. People with a temperature over 37.5 degrees have not been admitted inside the building. The number of non admitted people has been 132. In order to estimate the probability that a person has a temperature higher than 37.5

a) we can perform maximum likelihood inference of the rate λ od an exponential distribution, getting a point estimate $\hat{\lambda} = 2.349$

b) we can perform a maximum likelihood inference of the variance of a normal distribution, getting a point estimate $\hat{\sigma}^2 = 9.722$

c) we can perform maximum likelihood inference of the parameter p of a Bernoulli probability, getting a point estimate of $\hat{p} = 0.0363$

d) we can perform a maximum likelihood inference of the mean λ of a Poisson distribution, getting a point estimate $\hat{\lambda} = 1.434$

e) we can use the method of moments to infer the scale parameter β of a Gamma distribution, getting a point estimate $\hat{\sigma}^2 = 9.431$

Exercise 74. A game is such that the player independently in each round has a fixed (and unknown) probability p od winning. Suppose that the player plays several rounds of the game until he wins for the first time. He records in X_1 the number of rounds it took him to win for the first time. Suppose that the player repeats the whole sequence of rounds again and again (n times for total), each time recording in the variable X_i the number of attempts for the first win. At the end of the day, he has collected the sample $\mathbf{X} = \{X_1, \ldots, X_n\}$.

a) What would it be the parametric model (a parametric family of distributions) followed by the sample?

b) Prove that the beta family of distribution is conjugate to the family of distributions of the sample.

c) Another player, who tried the game before, tells you that, according to his/her experience, the probability of winning should be around 0.1, and most probably between 0 and 0.2. How would you encode this information in the Bayesian inferential procedure? What would it be the posterior distribution?

Exercise 75. An electrical scale gives a reading equal to the true weight plus a random error that is normally distributed with mean 0 and standard deviation $\sigma = 0.1$ mg. Suppose that the results of 5 successive weighings of the same object are as follows: 3.142, 3.163, 3.155, 3.150, 3.141.

a) Determine a 95% confidence interval estimate of the true weight.

b) Determine a 99% confidence interval estimate of the true weight.

Exercise 76. Suppose that when sampling from a normal population having an unknown mean μ and unknown variance σ^2, we wish to determine a sample size n so as to guarantee that the resulting $100(1-\alpha)\%$ confidence interval for μ will be of size no greater than A, for given values for α and A. Explain how we can approximately do this by a double sampling scheme that first takes a subsample of size 30 and then chooses the total sample size by using the results of the first subsample.

Exercise 77. A set of 10 determinations, by a method devised by the chemist Karl Fischer, of the percentage of water in a methanol solution yielded the following data:

$$0.50,\ 0.55,\ 0.53,\ 0.56,\ 0.54,\ 0.57,\ 0.52,\ 0.60,\ 0.55,\ 0.58$$

Assuming normality, use these data to give a 95% confidence interval for the actual percentage.

Exercise 78. Independent random samples are taken from the output of two machines on a production line. The weight of each item is of interest. From the first machine, a sample of size 36 is taken, with sample mean weight of 120g and a sample variance of 4. From the second machine, a sample of size 64 is taken, with a sample mean weight of 130g and a sample variance of 5. It is assumed that the weights of items from the first machine are normally distributed with mean μ_1 and variance σ_1^2 and that the weights of items from the second machine are normally distributed with mean μ_2 and variance σ_2^2. Find a 99% confidence interval for $\mu_1 - \mu_2$.

Exercise 79. Out of 100 randomly chosen individuals having lung cancer, 67 died within 5 years of detection.

a) Estimate the probability that a person contracting lung cancer will die within 5 years.

b) How large an additional sample would be required to be 95% confident that the error in estimating the probability in part (a) is less than 0.2?

Exercise 80. Suppose that lifetimes of batteries are exponentially distributed with mean θ. If the average of a sample of 10 batteries is 36 hours, determine a 95% two-sided confidence interval for θ.

Exercise 81. Consider two estimators d_1 and d_2 dof the parameter θ. If $E[d_1] = \theta$, $Var(d_1) = 6$ and $E[d_2] = \theta + 2, Var(d_2) = 2$, which estimator should be preferred?

Exercise 82. Suppose that the number of accidents occurring daily in a certain plant has a Poisson distribution with unknown mean λ. Based on previous experience in similar industrial plants, suppose that a statistician's initial feelings about the possible value of λ can be expressed by an exponential distribution with parameter 1. That is, the prior density is

$$p(\lambda) = e^{-\lambda}, 0 < \lambda < \infty \ .$$

Determine the Bayes estimate of λ if there are a total of 83 accidents over the next 10 days. What is the maximum likelihood estimate?

Exercise 83. A population distribution is known to have standard deviation 20. Determine the p-value of a test of hypothesis that the population mean is equal to 50, if the average of a sample of 64 observations is

a) 52.5

b) 55

c) 57.5

Exercise 84. The weights of salmon grown at a commercial hatchery are normally distributed with a standard deviation of 1.2 pounds. The hatchery claims that the mean weight of this year's crop is at least 7.6 pounds. Suppose a random sample of 16 fish yielded an average of 7.2 pounds. Is this strong enough evidence to reject the hatchery's claims at the

a) 5% level of significance;

b) 1% level of significance)

c) What is the p-value?

Exercise 85. There is some variability in the amount of phenobarbital in each capsule sold by a manifacturer. However, the manifacturer claims that the mean value is 20.0 mg. To test this, a sample of 25 pills yielded a sample mean of 19.7 with a sample standard deviation of 1.3. What inference would you draw from these data? In particular, are the data strong enough evidence to discredit the claim of the manifacturer? Use 5% level of significance.

Exercise 86. A car is advertised as having a gas mileage rating of at least 30 miles/gallon in highway driving. If the miles per gallon obtained in 10 independent experiments are 26, 24, 20, 25, 27, 25, 28, 30, 26, 33, should you believe the advertisement? What assumption are you making?

Exercise 87.A professor claims that the average starting salary of industrial engineering graduating seniors is greater than that of civil engineers graduates. To study this claim, samples of 16 industrial engineers and 16 civil engineers, all of whom graduated in 2006, were chosen and sample members were queried about their starting salaries. If the industrial engineers had a sample mean salary of \$59,700 and a sample standard deviation of \$2,400, and the civil engineers had a sample mean salary of \$58,400 and a sample standard deviation of \$2,200, has the professor's claim been verified? Find the appropriate p-value.

Exercise 88. An ambulance service claims that at least 45% of its calls involve life-threatening emergencies. To check this, a random sample of 200 calls was selected from the service's files. If 70 of these calls involved life-threatening emergencies, is the service's claim believable at the (a) 5% level of significance; (b) 1% level of significance?

Exercise 89. Among 100 vacuum tubes tested, 41 had lifetimes of less than 30 hours, 31 had lifetimes between 30 and 60 hours, 3 had lifetimes between 60 and 90 hours, and 15 had lifetimes of greater than 90 hours. Are these data consistent with the hypothesis that a vacuum tube's lifetime is exponentially distributed with a mean of 50 hours?

Exercise 90. The number of infant mortalities as function of the baby's birth weight (in grams) for 72,730 live white births in New York in 1974 as follows

Birthweight	**Alive**	**Dead**
Less than 2,500	4,597	618
Greater than 2,500	67,093	422

Test the hypothesis (with $\alpha = 0.05$) that the birth weight is independent of whether or not the baby survives its first year.

SOLUTIONS OF SOME OF THE EXERCISES OF THE FILE

"SMDS - Collection of exercises to be solved in the class"

Es 1

$S_{25} = X_1 + \cdots + X_{25}$ $\quad X_i : \quad E[X_i]=20, \quad V[X_i]=16$

$E[S_{25}] = 500 \quad V[S_{25}] = 400 \quad S_{25} \sim N(500, 400)$

$$P[S_n \leq 450] = P\left[\frac{S_{25}-500}{\sqrt{400}} \leq \frac{450-500}{\sqrt{400}}\right] = P[Z \leq -2.5]$$

$$= 1 - P[Z \leq 2.5] = 1 - \phi(2.5) = 0.006$$

Es 2

Sol:

$$X_i = \begin{cases} 1 & 1/6 \\ 2 & 1/6 \\ \vdots & \\ 6 & 1/6 \end{cases} \qquad E[X_i] = \frac{7}{2} \qquad V[X_i] = E[X_i^2] - (E[X_i])^2 = \frac{35}{12}$$

Since $S_{10} = X_1 + \cdots + X_{10}$ is discrete, then one can compute the following $P[29.5 \leq S_{10} \leq 40.5]$ using $S_{10} \sim N\left(\frac{70}{2}, \frac{350}{12}\right)$. Thus

$$P[29.5 \leq S_{10} \leq 40.5] = P\left[\frac{29.5-35}{\sqrt{\frac{350}{12}}} \leq Z \leq \frac{40.5-35}{\sqrt{\frac{350}{12}}}\right] = P[-1.02 \leq Z \leq 1.02]$$

$$= \phi(1.02) - (1 - \phi(1.02)) = 0.85 - (1 - 0.85) = 0.70$$

Es 3

Consider $\left(\prod_{i=1}^{n} X_i\right)^{1/n}$. Take logarithm ...

$$\log\left[\left(\prod_{i=1}^{n} X_i\right)^{1/n}\right] = \frac{1}{n}\sum_{i=1}^{n} \log(X_i) = \frac{Y_1 + \cdots + Y_n}{n} = \overline{Y}_n$$

where $Y_i = \log(X_i)$

By the LLN, $\bar{Y}_n \to E[Y_i] = E[\log(X_i)]$ Thus

$$\log\left[\left(\prod_{i=1}^{n} X_i\right)^{1/n}\right] \xrightarrow{p} E[\log(X_i)]$$

So that $$\left(\prod_{i=1}^{n} X_i\right)^{1/n} \xrightarrow{p} e^{E[\log(X_i)]}$$

Es 4

If the claim were true, then, by CLT,

$$\bar{X}_{100} = \frac{X_1 + \cdots + X_{100}}{100} \sim N(2.2, 0.0009)$$

$E[X_i] = 2.2$ $\quad V[X_i] = 0.3^2 = 0.09$ $\quad \bar{X}_{100} \sim N\left(E[X_i], \frac{V[X_i]}{100}\right)$

$$P[\bar{X}_{100} > 3.1] = P\left[Z > \frac{3.1 - 2.2}{\sqrt{0.0009}}\right] = P[Z > 5.196] \simeq 0$$

Es 5

$\mu = 20$ $\quad \sigma^2 = 20$. Let us apply the Markov Inequality to $Y = X - \mu$.

$E[Y] = 0$

$Var[Y] = Var[X - \mu] = Var[X] = \sigma^2 = 20$

$E[|Y|^2] = E[Y^2] = Var[Y] + E[Y]^2 = \sigma^2 + 0 = 20$

$$P[0 < X < 40] = P[|X - \mu| < 20]$$
$$= P[|Y| < 20]$$

$$\geq 1 - \frac{E[|Y|^2]}{20^2} = 1 - \frac{20}{20^2} = 19/20$$

<u>Note</u>: (Corollary of Markov Inequality)

$$\forall X: \quad P[|X-\mu|<a] \geq 1 - \frac{Var(X)}{a^2}$$

$\longrightarrow$ <u>Chebychev Inequality</u>

Es 6

X_i = lifetime bulb i $\quad E[X_i]=5 \quad V[X_i]=25$

$$S_{100} = X_1 + \dots + X_{100}$$

$$S_{100} \sim N(100 \cdot 5, 100 \cdot 25) = N(500, 2500)$$

$$P[S_{100} > 525] = P\left[Z > \frac{525-500}{\sqrt{2500}}\right] = 1 - \Phi(1/2)$$

Es 9

Recall that $\sum_{i=1}^{k} i = \frac{k(k+1)}{2} \geq \frac{k^2}{2}$. Thus

$$E[X] = \sum_{i=1}^{\infty} i\, P[X=i] \geq \sum_{i=1}^{k} i\, P[X=i]$$

$$\geq \sum_{i=1}^{k} i \cdot P[X=k] \quad \text{(since } P[X=i] \downarrow \text{ in } i\text{)}$$

$$\geq P[X=k] \cdot \frac{k(k+1)}{2} \geq P[X=k] \cdot \frac{k^2}{2}$$

Es 10

The first theoretical moment of Y is $\frac{\theta}{\theta+1}$:

$$E(Y) = \int_0^1 y \cdot \theta y^{\theta-1}\, dy$$
$$= \theta \cdot \frac{y^{\theta+1}}{\theta+1}\Big|_0^1$$
$$= \frac{\theta}{\theta+1}$$

Setting $E(Y)$ equal to $\frac{1}{n}\sum_{i=1}^{n} y_i (=\overline{y})$, the first sample moment, gives

$$\frac{\theta}{\theta+1} = \overline{y}$$

which implies that the method of moments estimate for θ is

$$\theta_e = \frac{\overline{y}}{1-\overline{y}}$$

Here, $\overline{y} = \frac{1}{4}(0.42+0.10+0.65+0.23) = 0.35$, so

$$\theta_e = \frac{0.35}{1-0.35} = 0.54$$

Es 11

$$L(\lambda) = e^{-\lambda}\frac{\lambda^3}{3!} \cdot e^{-\lambda}\frac{\lambda^5}{5!} \cdot e^{-\lambda}\frac{\lambda^4}{4!} \cdot e^{-\lambda}\frac{\lambda^2}{2!} = e^{-4\lambda}\lambda^{14}\frac{1}{3!5!4!2!}$$

Then $\ln L(\lambda) = -4\lambda + 14\ln\lambda - \ln(3!5!4!2!)$. Differentiating $\ln L(\lambda)$ with respect to λ gives

$$\frac{d\ln L(\lambda)}{d\lambda} = -4 + \frac{14}{\lambda}$$

To find the λ that maximizes $L(\lambda)$, we set the derivative equal to zero. Here $-4+\frac{14}{\lambda} = 0$ implies that $4\lambda = 14$, and the solution to this equation is $\lambda = \frac{14}{4} = 3.5$.

Es 12

$$L(\theta) = \prod_{i=1}^{n} \frac{1}{\theta^2} y_i e^{-y_i/\theta} \quad = \theta^{-2n} \left(\prod_{i=1}^{n} y_i \right) e^{-(1/\theta) \sum_{i=1}^{n} y_i}$$

Setting the derivative of ln $L(\theta)$ equal to 0 gives

$$\frac{d \ln L(\theta)}{d\theta} = \frac{-2n}{\theta} + \frac{1}{\theta^2} \sum_{i=1}^{n} y_i = 0$$

which implies that

$$\theta_e = \frac{1}{2n} \sum_{i=1}^{n} y_i$$

$$\sum_{i=1}^{5} y_i = 9.2 + 5.6 + 18.4 + 12.1 + 10.7 = 56.0,$$

$$\theta_e = \frac{1}{2(5)}(56.0) = 5.6$$

Es 13

MME) $\lambda = E[X] \longrightarrow \hat{\lambda} = \bar{X}_n \qquad (\bar{X}_n = M_1)$

MLE) Let $\bar{X} = (X_1, \ldots, X_n)$. Then

$$L(\lambda | \bar{x}) = \prod_{i=1}^{n} \frac{\lambda^{x_i}}{x_i!} e^{-\lambda} = \frac{\lambda^{\sum x_i} \cdot e^{-n\lambda}}{\prod_i x_i!}$$

$$\ln L(\lambda | \bar{x}) = -n\lambda + \sum_{i=1}^{n} x_i \ln(\lambda) - \ln\left(\prod x_i!\right)$$

$$0 = \frac{d \ln L(\lambda|\bar{x})}{d\theta} = -n + \frac{\sum_i x_i}{\lambda} \longrightarrow \hat{\lambda} = \frac{\sum x_i}{n} = \bar{X}_n$$

BAYES)

Assume the prior to be $\sim \Gamma(\alpha,\gamma)$, i.e.

$$\pi(\lambda) = \frac{\gamma e^{-\gamma\lambda}(\gamma\lambda)^{\alpha-1}}{\Gamma(\alpha)} = \frac{\gamma^{\alpha}\lambda^{\alpha-1}e^{-\gamma\lambda}}{\Gamma(\alpha)}$$

Let $S = X_1 + \dots + X_n$. $S \sim \text{Poiss}(n\lambda) \rightarrow f_S(s|\lambda) = \frac{(n\lambda)^s}{s!} e^{-n\lambda}$

Thus

$$\pi(\lambda|s) \sim f_S(s|\lambda)\cdot\pi(\lambda)$$

$$= \frac{(n\lambda)^s}{s!}e^{-n\lambda} \cdot \frac{\gamma^{\alpha}\lambda^{\alpha-1}e^{-\gamma\lambda}}{\Gamma(\alpha)}$$

$$= \underbrace{\frac{n^s}{s!}\frac{\gamma^{\alpha}}{\Gamma(\alpha)}}_{\text{constant in }\lambda} \cdot \underbrace{e^{-(n+\gamma)\lambda}\cdot\lambda^{s+\alpha-1}}_{\text{terms in }\lambda \rightsquigarrow \Gamma(s+\alpha,\, n+\gamma)}$$

The Bayes estimate of λ is $E[\Gamma(s+\alpha, n+\gamma)]$, i.e,

$$\hat{\lambda} = \frac{s+\alpha}{n+\gamma} \qquad (s = \textstyle\sum_i x_i) \qquad \left(\text{remember } E[\Gamma(\alpha,\gamma)] = \alpha/\gamma\right)$$

By experience $E[\pi] \cong 1.5$, $V[\pi] \cong 1$, since $E[\Gamma(\alpha,\gamma)] = \alpha/\gamma$ and $V[\Gamma(\alpha,\gamma)] = \alpha/\gamma^2$, then

$$\begin{cases} \alpha/\gamma = \frac{3}{2} \\ \alpha/\gamma^2 = 1 \end{cases} \Longrightarrow \alpha = 9/4 \quad \gamma = 3/2$$

Es 15

$$L(\theta|\bar{x}) = \theta^n \prod_{i=1}^{n} x_i^{\theta-1}$$

$$\ln L(\theta|\bar{x}) = n \cdot \ln\theta + (\theta-1)\ln\left(\prod x_i\right) = n\ln\theta + (\theta-1)\sum_i \ln(x_i)$$

$$\frac{d \ln(L(\theta|\bar{x}))}{d\theta} = \frac{n}{\theta} + \sum_i \ln(x_i) = 0 \qquad \hat{\theta}_n = -\frac{n}{\sum_i^n \ln(x_i)}$$

Es 16

$X \sim Bern(p)$. $\quad \pi(p) = \begin{cases} 1 & p \in (0,1) \\ 0 & p \notin (0,1) \end{cases}$

$$\pi(p|\bar{x}) = L(p|\bar{x}) \cdot \pi(p)$$

$$\sim p^{\Sigma x_i}(1-p)^{n-\Sigma x_i} \quad \mathbb{1}_{(0,1)}(p)$$

(a) $$\frac{\int_0^{0.2} p^2(1-p)^8\,dp}{\int_0^1 p^2(1-p)^8\,dp} = \frac{\left[-\frac{(1-p)^9}{9} + \frac{(1-p)^{10}}{5} - \frac{(1-p)^{11}}{11}\right]_0^{0.2}}{\left[-\frac{(1-p)^9}{9} + \frac{(1-p)^{10}}{5} - \frac{(1-p)^{11}}{11}\right]_0^{1}} = 0.61$$

Es 18

$$Y \sim U[\theta_1-\theta_2, \theta_1+\theta_2]$$

$$E[Y] = \frac{(\theta_1-\theta_2)+(\theta_1+\theta_2)}{2} = \theta_1$$

$$V[Y] = \frac{[(\theta_1+\theta_2)-(\theta_1-\theta_2)]^2}{12} = \frac{(2\theta_2)^2}{12} = \frac{\theta_2^2}{3}$$

$$E[Y^2] = V[Y] + (E[Y])^2 = \frac{\theta_2^2}{3} + \theta_1^2 = \frac{1}{3}\left(\theta_2^2 + 3\theta_1^2\right)$$

Solve $\begin{cases} M_1 = \hat{\theta}_1 \\ M_2 = \frac{1}{3}\left(\hat{\theta}_2^2 + 3\hat{\theta}_1^2\right) \end{cases}$

where
$$M_1 = \frac{8.3 + 4.9 + 2.6 + 6.5}{4} = 5.575$$
$$M_2 = \frac{(8.3)^2 + (4.9)^2 + (2.6)^2 + (6.5)^2}{4} = 35.4775$$

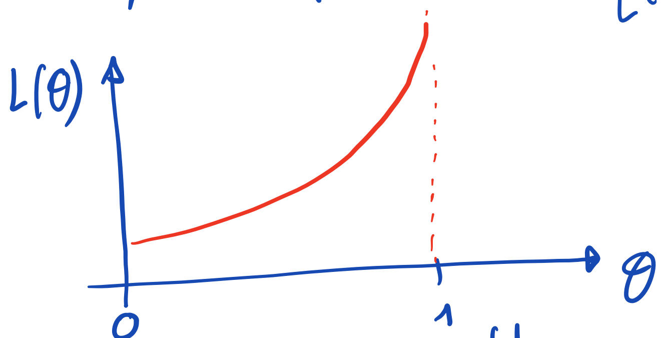
$$\Rightarrow \quad \hat{\theta}_1 = 5.575 \qquad \hat{\theta}_2 = \sqrt{3 \cdot M_2 - 3 \cdot \hat{\theta}_1^2} \cong 3.63$$

Es 26

Answer : $\hat{\theta} = \min\{y_i\} = 0.21$

You must consider it as a function of θ, with the y_i fixed constant.

Consider only the first part $\left[\frac{2^n}{(1-\theta^2)^n} \cdot \prod_{i=1}^{n} y_i\right]$



It is increasing, thus the max is in 1

But... could θ be greather than the y_i ?

Consider $\prod_{i=1}^{n} \mathbb{1}_{[\theta,1]}(y_i)$

If $\theta > y_i$ for at least one y_i (say, e.g, if $\theta > y_1$) then $\mathbb{1}_{[\theta,1]}(y_1) = 0$, and the product becomes 0.

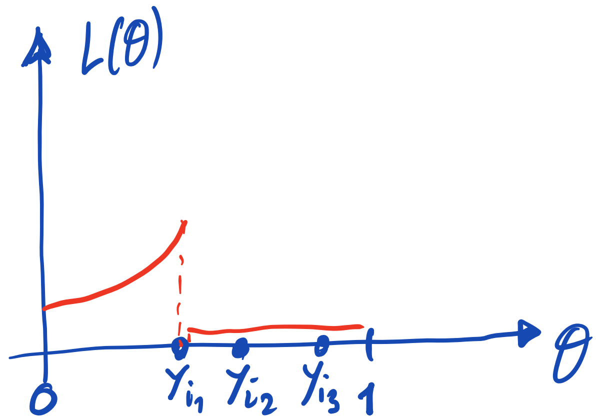
Thus θ must be smaller than all the y_i, i.e., it must be

$$\theta \leq y_i \quad \forall i$$

i.e.,

$$\theta \leq \min(y_i)$$

Because of it, the correct graph of $L(\theta)$ is



and the $\max L(\theta)$ is attained in $\min\{y_i\}$.

$$\hat{\theta} = \text{MLE} = \arg\max L(\theta) = \min\{y_i\}$$

Es 28

a)

$$p_{\hat{p}}\left(\frac{k}{10}\right) = P\left(\hat{p} = \frac{k}{10}\right) = P(X = k) = \binom{10}{k}(0.60)^k(0.40)^{10-k}, \quad k = 0, 1, \ldots, 10$$

Therefore,

$$\begin{aligned} P\left(\left|\frac{X}{10} - 0.60\right| \le 0.10\right) &= P\left(0.60 - 0.10 \le \frac{X}{10} \le 0.60 + 0.10\right) \\ &= P(5 \le X \le 7) \\ &= \sum_{k=5}^{7}\binom{10}{k}(0.60)^k(0.40)^{10-k} \end{aligned}$$

b)

$$\begin{aligned} P\left(\left|\frac{X}{100} - 0.60\right| \le 0.10\right) &= P\left(0.50 \le \frac{X}{100} \le 0.70\right) \\ &= P\left[\frac{0.50 - 0.60}{\sqrt{\frac{(0.60)(0.40)}{100}}} \le \frac{X/100 - 0.60}{\sqrt{\frac{(0.60)(0.40)}{100}}} \le \frac{0.70 - 0.60}{\sqrt{\frac{(0.60)(0.40)}{100}}}\right] \\ &\doteq P(-2.04 \le Z \le 2.04) \\ &= 0.9586 \end{aligned}$$

Es 29

$$E(\hat{\theta}_1) = E\left(\frac{3}{2}\overline{Y}\right) = \frac{3}{2}E(\overline{Y}) = \frac{3}{2}E(Y) = \frac{3}{2}\cdot\frac{2}{3}\theta = \theta$$

$$f_{Y_{\max}}(y) = nF_Y(y)^{n-1}f_Y(y)$$

The cdf for Y is

$$F_Y(y) = \int_0^y \frac{2t}{\theta^2}dt = \frac{y^2}{\theta^2}$$

Then

$$f_{Y_{\max}}(y) = n\left(\frac{y^2}{\theta^2}\right)^{n-1}\frac{2y}{\theta^2} = \frac{2n}{\theta^{2n}}y^{2n-1}, 0 \le y \le \theta$$

Therefore,

$$E(Y_{\max}) = \int_0^\theta y \cdot \frac{2n}{\theta^{2n}}y^{2n-1}dy = \frac{2n}{\theta^{2n}}\int_0^\theta y^{2n}dy = \frac{2n}{\theta^{2n}} \cdot \frac{\theta^{2n+1}}{2n+1} = \frac{2n}{2n+1}\theta$$

$\lim\limits_{n\to\infty} \frac{2n}{2n+1}\theta = \theta$. Intuitively, this decrease in the bias makes sense because f_{θ_2} becomes increasingly concentrated around θ as n grows. ■

Es 30

$$E(Y^2) = \int_0^\theta y^2 \cdot \frac{2y}{\theta^2}dy = \frac{2}{\theta^2}\int_0^\theta y^3 dy = \frac{2}{\theta^2} \cdot \frac{\theta^4}{4} = \frac{1}{2}\theta^2$$

and

$$\text{Var}(Y) = E(Y^2) - E(Y)^2 = \frac{1}{2}\theta^2 - \left(\frac{2}{3}\theta\right)^2 = \frac{\theta^2}{18}$$

Then

$$\text{Var}(\hat{\theta}_1) = \text{Var}\left(\frac{3}{2}\overline{Y}\right) = \frac{9}{4}\text{Var}(\overline{Y}) = \frac{9}{4}\frac{\text{Var}(Y)}{n} = \frac{9}{4n} \cdot \frac{\theta^2}{18} = \frac{\theta^2}{8n}$$

To address the variance of $\hat{\theta}_2 = \frac{2n+1}{2n}Y_{\max}$, we start with finding the variance of $Y_{\max}$. Recall that its pdf is

$$nF_Y(y)^{n-1}f_Y(y) = \frac{2n}{\theta^{2n}}y^{2n-1}, 0 \le y \le \theta$$

From that expression, we obtain

$$E(Y_{\max}^2) = \int_0^\theta y^2 \cdot \frac{2n}{\theta^{2n}}y^{2n-1}dy = \frac{2n}{\theta^{2n}}\int_0^\theta y^{2n+1}dy = \frac{2n}{\theta^{2n}} \cdot \frac{\theta^{2n+2}}{2n+2} = \frac{n}{n+1}\theta^2$$

and then

$$\text{Var}(Y_{\max}) = E(Y_{\max}^2) - E(Y_{\max})^2 = \frac{n}{n+1}\theta^2 - \left(\frac{2n}{2n+1}\theta\right)^2 = \frac{n}{(n+1)(2n+1)^2}\theta^2$$

Finally,

$$\text{Var}(\hat{\theta}_2) = \text{Var}\left(\frac{2n+1}{2n}Y_{\max}\right) = \frac{(2n+1)^2}{4n^2}\text{Var}(Y_{\max}) = \frac{(2n+1)^2}{4n^2} \cdot \frac{n}{(n+1)(2n+1)^2}\theta^2$$

$$= \frac{1}{4n(n+1)}\theta^2$$

Es 31

Note, first, that

$$\text{Var}(\hat{p}) = \text{Var}\left(\frac{X}{n}\right) = \frac{1}{n^2}\text{Var}(X) = \frac{1}{n^2}np(1-p) = \frac{p(1-p)}{n}$$

(since X is a binomial random variable). To evaluate, say, the *second* form of the Cramér-Rao lower bound, we begin by writing

$$\ln p_{X_i}(X_i; p) = X_i \ln p + (1 - X_i)\ln(1-p)$$

Moreover,

$$\frac{\partial \ln p_{X_i}(X_i; p)}{\partial p} = \frac{X_i}{p} - \frac{1 - X_i}{1-p}$$

and

$$\frac{\partial^2 \ln p_{X_i}(X_i; p)}{\partial p^2} = -\frac{X_i}{p^2} - \frac{1 - X_i}{(1-p)^2}$$

Taking the expected value of the second derivative gives

$$E\left[\frac{\partial^2 \ln p_{X_i}(X_i; p)}{\partial p^2}\right] = -\frac{p}{p^2} - \frac{(1-p)}{(1-p)^2} = -\frac{1}{p(1-p)}$$

The Cramér-Rao lower bound, then, reduces to

$$\frac{1}{-n\left[-\frac{1}{p(1-p)}\right]} = \frac{p(1-p)}{n}$$

which *equals* the variance of $\hat{p} = \frac{X}{n}$. It follows that $\frac{X}{n}$ is the preferred statistic for estimating the binomial parameter p: No unbiased estimator can possibly be more precise.

Es 32

$$f_{Y_{\max}}(y) = \frac{ny^{n-1}}{\theta^n}, \quad 0 \le y \le \theta$$

Therefore,

$$P(|\hat{\theta}_n - \theta| < \varepsilon) = P(\theta - \varepsilon < \hat{\theta}_n < \theta) = \int_{\theta-\varepsilon}^{\theta} \frac{ny^{n-1}}{\theta^n}\,dy = \left.\frac{y^n}{\theta^n}\right|_{\theta-\varepsilon}^{\theta}$$
$$= 1 - \left(\frac{\theta - \varepsilon}{\theta}\right)^n$$

Since $[(\theta - \varepsilon)/\theta] < 1$, it follows that $[(\theta-\varepsilon)/\theta]^n \to 0$ as $n \to \infty$. Therefore, $\lim_{n\to\infty} P(|\hat{\theta}_n - \theta| < \varepsilon) = 1$, proving that $\hat{\theta}_n = Y_{\max}$ is consistent for θ.

Es 42

- Small sample, unknown μ and σ^2

$$\left(\bar{X}_n - t_{n-1,1-\alpha/2}\sqrt{\frac{\hat{\sigma}_n^2}{n}},\ \bar{X}_n + t_{n-1,1-\alpha/2}\sqrt{\frac{\hat{\sigma}_n^2}{n}}\right)$$

$\bar{X}_n = 80.57$ $\quad \hat{\sigma}_n^2 = 49.8$ $\quad n=6$ $\quad \alpha = 0.1$

$$t_{5,\,0.95} = 2.01$$

$$\left(80.57 - 2.01\cdot\sqrt{\frac{49.8}{6}},\ 80.57 + 2.01\cdot\sqrt{\frac{49.8}{6}}\right) \simeq (74.9\,;\ 86.8)$$

Es 42

$$e/2 = 3 = \underbrace{Z_{0.84}}_{\substack{\downarrow \\ \text{table} \\ \downarrow \\ 1.00}} \cdot \sqrt{\frac{68}{n}}$$

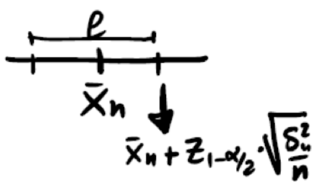
$1-\alpha = 0.68$

$\downarrow$

$\alpha = 0.32$

$\alpha/2 = 0.16$

$Z_{1-\alpha/2} = Z_{0.84}$



Thus

$$3 = 1\cdot\sqrt{\frac{68}{n}} \longrightarrow \sqrt{n} = \frac{\sqrt{68}}{3} \qquad n = \frac{68}{9} = 7.5 \longrightarrow n \geq 8$$

Es 48

Sol:

The difference $X_{n+1} - \bar{X}$ follows the Gaussian distribution with vanishing mean and a variance of $\sigma^2(1 + \frac{1}{n})$. If sigma is known we can conclude that

$$P\left(-z_{\frac{\alpha}{2}} < \frac{X_{n+1} - \bar{X}}{\sigma\sqrt{1 + \frac{1}{n}}} < z_{\frac{\alpha}{2}}\right) = 1 - \alpha$$

Inverting this region we get the following $(1 - \alpha)$-prevision interval

$$\left[\bar{X} - z_{\frac{\alpha}{2}}\sigma\sqrt{1 + \frac{1}{n}}, \bar{X} + z_{\frac{\alpha}{2}}\sigma\sqrt{1 + \frac{1}{n}}\right]$$

In case the variance is unknown, we can proceed by combining the following two arguments.

1. Let S_n^2 be the estimator $S_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$, then

$$(n-1)\frac{S_n^2}{\sigma^2} \sim \chi_{n-1}^2$$

2. The Student's t-distribution with n degrees of freedom can be defined as the distribution of the following ratio

$$T = \frac{Z}{\sqrt{Y/n}}$$

where Z and Y are independent. Z is a standard normal with expected value 0 and variance 1; Y has a chi-squared distribution with n degrees of freedom.

Putting the two things together we obtain

$$\frac{\frac{X_{n+1}-\bar{X}}{\sigma\sqrt{1+\frac{1}{n}}}}{\sqrt{\frac{(n-1)S_n^2}{(n-1)\sigma^2}}} = \frac{X_{n+1}-\bar{X}}{S_n\sqrt{1+\frac{1}{n}}} \sim t_{n-1}$$

and therefore we can easily get that the interval

$$\left[\bar{X} - t_{\frac{\alpha}{2},n-1}\, S_n\sqrt{1+\frac{1}{n}}, \bar{X} + t_{\frac{\alpha}{2},n-1}\, S_n\sqrt{1+\frac{1}{n}}\right]$$

is a $(1-\alpha)$-prevision interval for X_{n+1}.

a) $I = \left(\bar{X}_n - z_{0,995}\sqrt{\frac{S_n^2}{n}} ; \bar{X}_n + z_{0.995}\sqrt{\frac{S_n^2}{n}}\right)$

$= \left(12.5 - 2.58\sqrt{\frac{0,36}{50}} ; 12,5 + 2,58\sqrt{\frac{0,36}{50}}\right)$

$= (12.28 ; 12.72)$

b) $Z = \frac{12.5 - 12.00}{\sqrt{\frac{S_n^2}{n}}} = 5.89$

$C = (-z_{0.975}, +z_{0.975}) = (-1,96 ; +1.96)$

$Z \notin C \Rightarrow$ reject H_0

Es 54

$$\bar{X}_5 = \frac{12.5 + \dots + 13.1}{5} = 12.4 \qquad S_5^2 = \frac{(12.5-12.4)^2 + \dots + (13.1-12.4)^2}{4} = 0.86$$

a) $I = \left(\bar{X}_n - t_{n-1,1-\alpha/2} \cdot \sqrt{\frac{S_n^2}{n}} \,;\, \bar{X}_n + t_{n-1,1-\alpha/2}\sqrt{\frac{S_n^2}{n}}\right)$

$= \left(12.4 - 4.6 \cdot \sqrt{\frac{0.86}{5}} \,;\, 12.4 + 4.6 \cdot \sqrt{\frac{0.86}{5}}\right) \qquad t_{4,0.995} = 4,6$

$= (10.49 \,;\, 14.41)$

b) $T = \frac{\bar{X}_n - \mu_0}{\sqrt{\frac{S_n^2}{n}}} = \frac{12.4 - 12}{\sqrt{\frac{0.86}{5}}} = 0.96$

$C = (-t_{n-1,1-\alpha/2} \,;\, +t_{n-1;1+\alpha/2}) = (-t_{4,0.975}, +t_{4,0.975})$

$= (-2.78 \,;\, +2.78)$

$T \in C \Rightarrow$ Do not reject H_0

Es 55

$$\bar{X}_5 = \frac{12.5 + \dots + 13.1}{5} = 12.4$$

$$S_5^2 = \frac{(12.5-12.4)^2 + \dots + (13.1-12.4)^2}{4} = 0.86$$

a) $I = \left(\frac{(n-1)S_n^2}{\chi^2_{n-1,1-\alpha/2}} , \frac{(n-1)S_n^2}{\chi^2_{n-1,\alpha/2}}\right) \qquad \chi^2_{4,0.975} = 11.14$

$\chi^2_{4,0.025} = 0.48$

$= \left(\frac{4 \cdot 0.86}{11.14} \,;\, \frac{4 \cdot 0.86}{0.48}\right) = (0.31 \,;\, 7.17)$

b) $V = \frac{(n-1) \cdot S_n^2}{\sigma_0^2} = \frac{4 \cdot 0.86}{0.5} = 6.88$

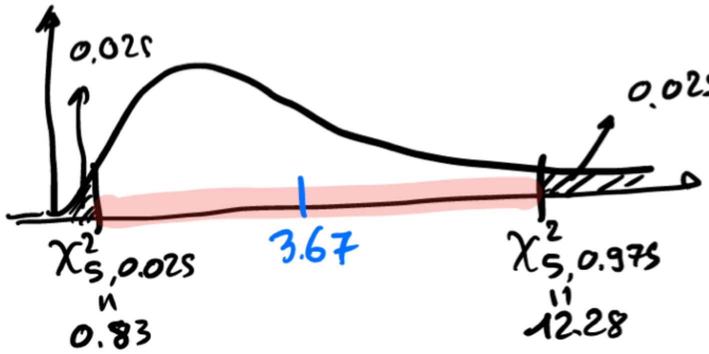
$C = (\chi^2_{n-1,\alpha/2} , \chi^2_{n-1,1-\alpha/2}) = (0.48 \,;\, 11.14)$

$V \in C \Rightarrow$ Do not reject H_0

Es 56

To test variance $(H_0: \sigma^2 = 2)$ we must use $V = \frac{(n-1) S_n}{\sigma_0} = \frac{5 \cdot 1.47}{2} = 3.67$

The acceptance region is determined by a χ^2_{n-1} ($n-1=5$ degrees freedom)

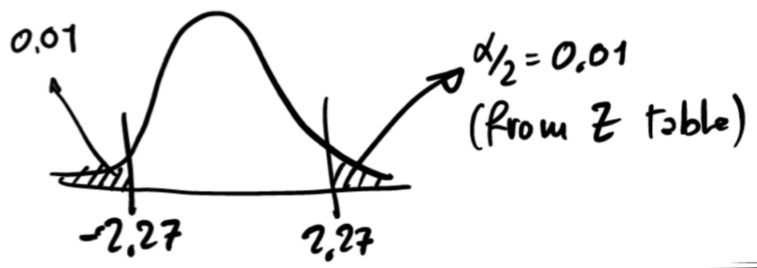


Since $V \in$ acept.
then we do not reject $H_0: \sigma^2 = 2$

Here acceptance region $= [-1, 1]$ (for $\bar{X}_n$)

That is, we accept if $\bar{X}_n \in [-1, 1]$

which means if $Z = \frac{\bar{X}_n - 0}{\sqrt{2/n}} \in \left[\frac{-1}{\sqrt{2/10}}, \frac{+1}{\sqrt{2/10}}\right] = [-2.27, +2.27]$



Thus $\alpha = 0.02$
Thus confidence 0.98

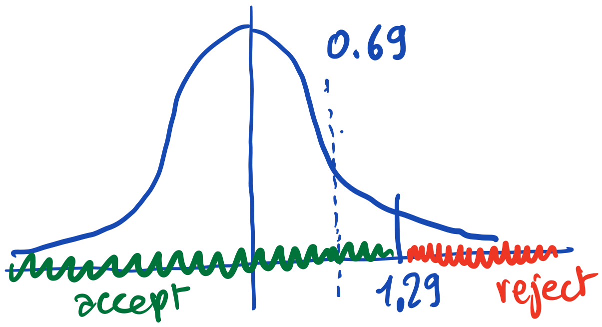
Es 57

$X_i = \begin{cases} 1 & p \\ 0 & 1-p \end{cases}$ (if spoiled)

$\bar{X}_n \sim N\left(p, \frac{p(1-p)}{n}\right)$

$H_0: p \leq 0.15$ $\quad H_1: p > 0.15$ (we check if there is evidente to reject H_0)

$\bar{X}_n = \frac{\frac{16}{91} - 0.15}{\sqrt{\frac{0.15(1-0.15)}{91}}} \simeq 0.69$

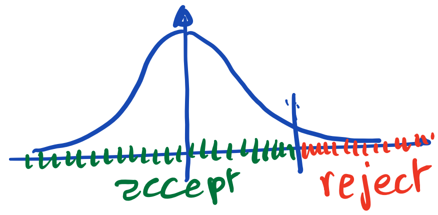


We can not reject $p \leq 15$ (more than 15% are contaminate

Es 58

Note: this is a one-sided test $\quad H_0: \mu_{POLITO} \leq 34.30$

(We must check if there is evidence to reject this claim)



It seems, because $\mu_{POLI} - \mu_0 \simeq 5$ which seems large

$n = 26 \qquad \bar{X}_n = 39.54 \qquad \hat{\sigma}_n = 5.2 \qquad \hat{\sigma}_n^2 = 5.2^2$

$$T = \frac{\bar{X}_n - \mu_0}{\sqrt{\hat{\sigma}_n / n}} = 5.138 \qquad \text{with 25 d.f.}$$

$\alpha = 0.05 \longrightarrow t_{25,0.95} = 1.71 \longrightarrow$ reject $\mu_{POLITO} \leq 34.3$

$\alpha = 0.01 \longrightarrow t_{25,0.99} = 2.49 \longrightarrow$ reject "

p-value

$$p = P[T > 5.138] = 1 - 0.999987 \simeq 1 \cdot 10^{-5}$$

$p < \alpha$ (in both cases)

$\longrightarrow$ We reject (both cases)

Es 59

Let $p_0 = 2.7 \cdot 10^{-4}$ be the proportion in population (whole)

p_1 = proportion in the specific region

$$X_i = \begin{cases} 1 & \rightsquigarrow \text{prob } p \\ 0 & \rightsquigarrow \text{prob } 1-p \end{cases}$$

H_0: $p_1 = p_0$ but unilateral test

(we reject if $\hat{p}_1$ is too "big")

Use $\alpha = 0.05$ $\qquad$ $n = 42\,500$

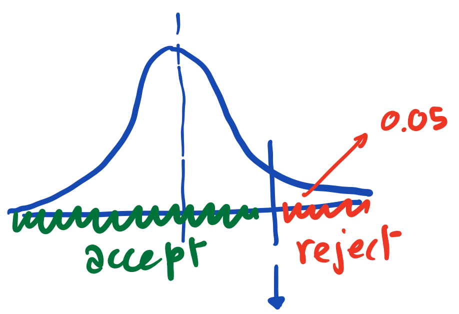
① By CLT, since n is large

$$S_n = \sum^{42500} X_i \sim N(n \cdot p_0,\ n p_0 (1-p_0))$$

(if $p_1 = p_0$)

so that $Z = \dfrac{S_n - E[S_n]}{\sqrt{V[S_n]}} \sim N(0,1)$

The observed Z is $\qquad Z_{obs} = \dfrac{14 - E[S_n]}{\sqrt{V[S_n]}} = 0.74$



$z_{0.95} = 1.65$

$Z_{obs} \in$ accept region

$\Downarrow$

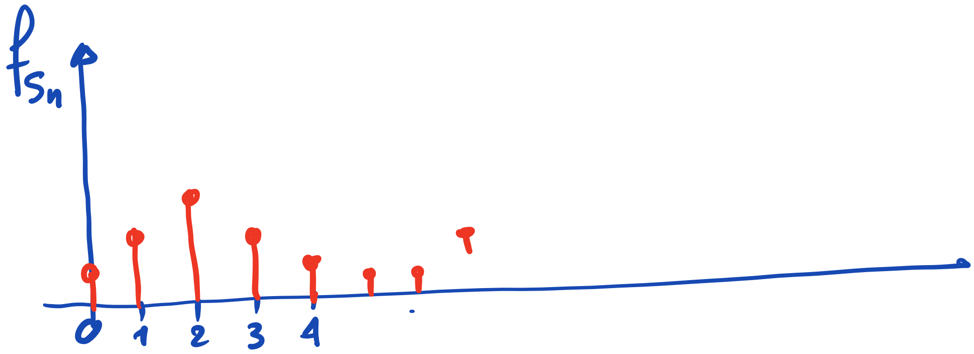
We do not reject

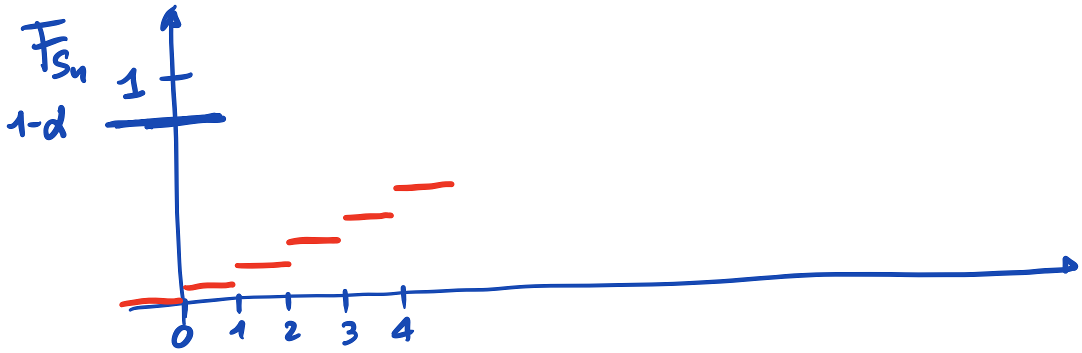
(Note: $Z > 1.65 \Rightarrow \dfrac{S_n - E[S_n]}{\sqrt{V[S_n]}} > 1.65 \Rightarrow S_n > 17.06$)

② Actually, there is no need of CLT.

In fact,

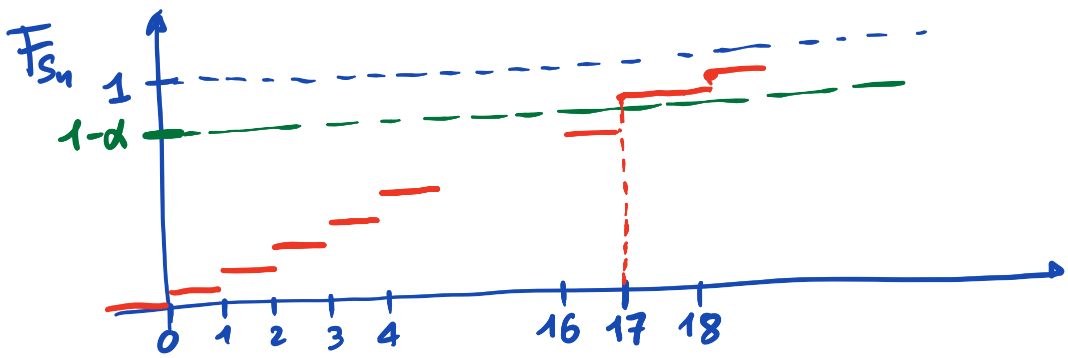
$$S_n \sim Bin(42'500,\ p_0) \qquad \text{(if } H_0 \text{ true)}$$





$\alpha = P[\text{error I kind}] = P[S_n > c \mid p_1 = p_0]$

thus c is the quantile of level $1-\alpha$
from a $\text{Bin}(42500, p_0)$



Such a c is 17 (long calculations)

Thus, for S_n $\rightarrow$ acceptance region $[0, 16]$

$\rightarrow$ rejection region $[17, +\infty)$

We observed 14. $\Rightarrow$ We do not reject.

Es 62

value	f_{ai}	p_i	f_{ti}
0	5	0,36	7,2
1	7	0,36	7,2
2	3	0,18	3,6
>2	5	0,10	2,0

$P_0 = \frac{1^0}{0} e^{-1} =$

$P_1 = \frac{1^1}{1!} e^{-1} =$

$\vdots$

$P_{>2} = 1 - P_0 - P_1 - P_2$

$$W = \sum_i \frac{(f_{ai} - f_{ti})^2}{f_{ti}} = 5,28$$

$$\chi^2_{N-m-1} \rightarrow \chi^2_{4-0-1} = \chi^2_3 \qquad q_{0.95} = 7.8$$

$W < q_{0.95} \longrightarrow$ accept H_0

Es 63

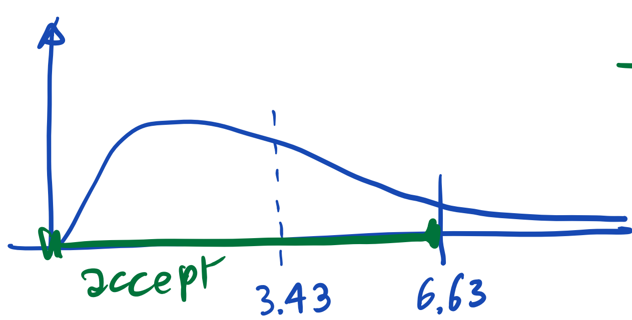
Y \ X	0	1	
A	10	40	50
B	15	135	150
	25	175	200

$\longrightarrow$

Y \ X	0	1	
A	0.05	0.2	0.25
B	0.075	0.675	0.75
	0.125	0.875	1

$$W = \left[\frac{(10 - 0.25 \cdot 0.125 \cdot 200)^2}{0.25 \cdot 0.125 \cdot 200} + \cdots \right] \cong 3.43$$

$$\chi^2_{(2-1)(2-1),\, 1-0.01} = \chi^2_{1,\,0.99} = 6.63$$



$\rightarrow$ We can not reject

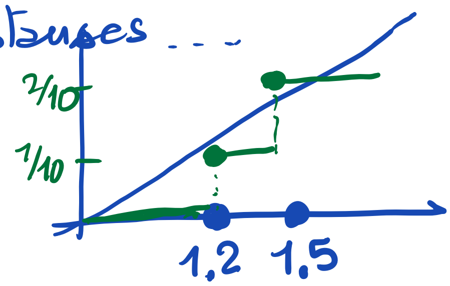
Es 64

Here use K-S test for GOF. $F(x) = \frac{x}{10}$ $(x \in [0, 10])$

For any x_i we consider 2 distances ...

$d_i^+ = \hat{F}(x_i) - F(x_i^-)$

$d_i^- = F(x_i) - \hat{F}(x_i^+)$



$d_1^+ = 0.12 - 0 = 0.12$ $d_1^- = 0,12 - 0.1 = 0.02$

$d_2^+ = 0.15 - 0,1 = 0.05$ $d_2^- = 0.15 - 0.2 = -0,05$

$d_3^+ = 0,23 - 0.2 = 0,03$ $d_3^- = -0,07$

$d_4^+ = 0,24 - 0,3 = -0.06$ $d_4^- = -0,16$

$d_5^+ = 0.4 - 0.4 = 0$ $d_5^- = -0.1$

$d_6^+ = 0.55 - 0.5 = 0.05$ $d_6^- = -0.05$

$d_7^+ = 0.61 - 0.6 = 0.01$ $d_7^- = -0,09$

$d_8^+ = 0.78 - 0,7 = 0.08$ $d_8^- = -0,02$

$d_9^+ = 0,81 - 0,8 = 0.01$ $d_9^- = -0.09$

$d_{10}^+ = 0.95 - 0,9 = 0.05$ $d_{10}^- = -0,05$

$D = \max\{|d_i^+|, |d_i^-|\} = 0,16$

$D_{10, 0.95} = 0,41$ $C = (0,41, 1]$

$D \in \bar{C} \Rightarrow$ We can not reject H_0

71

Quiz [redacted]. Dopo un ampio studio clinico (1500 pazienti) è stato stimato (nel senso di stima puntuale) che la probabilità di sviluppare un effetto avverso specifico è 0,0173. Usando l'approssimazione normale, quale delle seguenti e' corretta?

a) un intervallo di confidenza unilaterale del 95% per p è $(-\infty, 0,0482)$;

b) un intervallo di confidenza unilaterale del 95% per p è $(-\infty, 0,0228)$;

c) un intervallo di confidenza bilaterale del 95% per p è $(0,0023, 0,0273)$;

d) nessuna delle altre opzioni è corretta;

e) un intervallo di confidenza bilaterale del 95% per p è $(0,0168, 0,0178)$.

$\bar{X}_n \sim N\left(\hat{p}, \frac{\hat{p}(1-\hat{p})}{n}\right)$ $\qquad \hat{p} = 0.0173$ $\qquad z_{0.95} = 1.65$

$z_{0.975} = 1.96$

$I_{uni} = \left(-\infty, \hat{p} + 1.65 \cdot \sqrt{\frac{p(1-p)}{n}}\right) = (-\infty, 0.0228)$

$I_{bil} \left(\hat{p} - 1.96 \cdot \sqrt{\frac{p(1-p)}{n}}, \hat{p} + 1.96 \cdot \sqrt{\frac{p(1-p)}{n}}\right) = (0.0107, 0.0238)$

72

Quiz [redacted]. Se si desidera ridurre l'ampiezza attesa dell'intervallo di confidenza per la media di un campione normale di un fattore di 2 ...

a) ... è necessario ridurre la dimensione del campione di un fattore 2

b) ... è necessario raccogliere un campione con una dimensione del campione 4 volte superiore a quella attuale

c) ... nessuna delle altre opzioni è corretta

d) ... è necessario raccogliere un campione con una doppia dimensione del campione

e) ... è necessario raccogliere un campione con una dimensione del campione 4 volte inferiore a quella attuale

Amplitude $= 2 \cdot z_{1-\alpha/2} \cdot \sqrt{\frac{\sigma^2}{n}}$ $\rightarrow$ to half it use $4n$

Es 74

a) $X \sim Geo(p)$ $\qquad P[X=k] = p(1-p)^{k-1}$

b) (We have seen it already for $\bar{x} = x_1$)

$\pi = Beta(r,s)$ $\qquad \pi_\Theta(\theta) = \frac{\Gamma(r,s)}{\Gamma(r)\Gamma(s)} \theta^{r-1}(1-\theta)^{s-1}$ $\qquad \theta \equiv p$

$\pi_\Theta(\theta|\bar{x}) \sim L(\bar{x}|\theta)\,\pi_\Theta(\theta)$ $\qquad L(\bar{x}|\theta) = \theta^n(1-\theta)^{\Sigma x_i - n}$

$$\sim \theta^n(1-\theta)^{\Sigma x_i - n} \cdot \frac{\Gamma(r,s)}{\Gamma(r)\Gamma(s)} \theta^{r-1}(1-\theta)^{s-1}$$

$$\sim \underbrace{\theta^{n+r-1}(1-\theta)^{\Sigma x_i + s - n - 1}}$$

$\hookrightarrow$ variable part of a $Beta(n+r, \Sigma x_i + s - n)$

c) $E[\pi] = \frac{r}{r+s} \cong 0.1$ $\qquad V[\pi] = \frac{r \cdot s}{(r+s)^2(r+s+1)} \cong 0.01$

$\longrightarrow$ For example $r=1$, $s=10$. In this case

$\pi_\Theta(\theta|\bar{x}) \sim Beta(n+1, \Sigma x_i + 9)$

Es 75

$X_i = \mu + E_i$, $E_i \sim N(0, 0.01)$

$X_i \sim N(\mu, 0.01)$ $\qquad \bar{X}_5 = 3.150$ $\qquad \sigma = 0.01$

Use Z because σ^2 is know

$Z_{1-\alpha/2}$ $\qquad Z_{0.975} = 1.96$ $\qquad Z_{0.995} = 2.58$

$$I_{95\%} = \left(3.150 - 1.96\sqrt{\frac{0.01}{5}},\ 3.150 + 1.96\sqrt{\frac{0.01}{5}}\right) = (3.106, 3.194)$$

$$I_{99\%} = \left(3.150 - 2.58\sqrt{\frac{0.01}{5}},\ 3.150 + 2.58\sqrt{\frac{0.01}{5}}\right) = (3.093, 3.208)$$

Es 76

- First estimate σ^2: $\hat{\sigma}^2_{30} = \frac{\sum (x_i - \bar{X}_{30})^2}{29}$
- Now: $A/2 \leq z_{1-\alpha/2} \cdot \sqrt{\frac{\hat{\sigma}^2_{30}}{n}}$

$$\rightarrow \sqrt{n} \geq \frac{z_{1-\alpha/2}}{A/2} \cdot \sqrt{\hat{\sigma}^2_{30}}$$

$$n \geq \left(\frac{2 \cdot z_{1-\alpha/2}}{A}\right)^2 \cdot \hat{\sigma}^2_{30}$$

Es 77

$\bar{X}_{10} = 0.55$ $\qquad \hat{\sigma}^2_{10} = 0.000867$ $\qquad t_{9,\,1-\alpha/2} = 2.26$

$$I_{95\%} = \left(0.55 - 2.26\sqrt{\frac{\hat{\sigma}^2_{10}}{10}},\ 0.55 + 2.26\sqrt{\frac{\hat{\sigma}^2_{10}}{10}}\right)$$

$$= (0.529,\ 0.571)$$

(small sample, σ^2 unknown)

Es 78

$A \rightarrow \bar{X}^A_{36} = 120$, $\hat{\sigma}^{2\,A}_{30} = 4$ / $B \rightarrow \bar{X}^B_{64} = 130$ $\quad \hat{\sigma}^{2\,B}_{64} = 5$

$\bar{X}^A - \bar{X}^B = -10$ $\qquad z_{1-\alpha/2} = z_{0.995} = 2.58$

$$I_{99\%} = \left(-10 - 2.58 \cdot \sqrt{\frac{4}{36} + \frac{5}{64}},\ -10 + 2.58 \cdot \sqrt{\frac{4}{36} + \frac{5}{64}}\right) = \ldots$$

(large samples...)

ES 79

a) $\hat{p} = \frac{67}{100} = 0.67 = \bar{X}_n$

b) $\bar{X}_n \sim N\left(p, \frac{p(1-p)}{n}\right) \longrightarrow \hat{\sigma}^2_{\bar{X}_n} \sim \frac{0.67(0.33)}{n} = \frac{0.22}{n}$

$$A = 2 \cdot z_{0.975} \cdot \sqrt{\frac{0.22}{n}} < 0.02$$

$$\longrightarrow \quad \sqrt{n} > \frac{2 \cdot z_{0.975} \cdot \sqrt{0.22}}{0.02} = 91.9$$

$$n > (92)^2$$

ES 80

(small sample)

- We use the second method described in lectures.

$$I_{(1-\alpha)\%} = \left[\frac{\chi^2_{2n,\alpha/2}}{2\sum X_i} , \frac{\chi^2_{2n,1-\alpha/2}}{2\sum X_i} \right] = [0.0133 , 0.0475] \ (\text{for } \lambda)$$

$\left(\sum X_i = 360 \qquad \chi^2_{20,0.025} = 9.59 \qquad \chi^2_{20,0.975} = 34.17 \right)$

thus, for the mean, $I_{95\%} = (21.07 , 75.08)$

Es 81

$$MSE(d_1) = Var(d_1) + Bias(d_1)^2 = 6 + 0^2 = 6$$

$$MSE(d_2) = Var(d_2) + Bias(d_2)^2 = 2 + 2^2 = 6$$

Same MSE .... I would use the unbiased one..

Es 82

prior: $\pi(\lambda) = e^{-\lambda}$

$$L(\lambda|\bar{x}) = \prod_{i=1}^{n}\left(\frac{\lambda^{x_i}}{x_i!}e^{-\lambda}\right) = K \cdot \lambda^{\Sigma x_i} \cdot e^{-n\lambda}$$

Thus the posterior is proportional to $L(\lambda|\bar{x}) \cdot \pi(\lambda)$

$$\pi(\lambda|\bar{x}) \sim K \cdot \underbrace{\lambda^{\Sigma x_i} e^{-(n+1)\lambda}}$$

this is $\sim \Gamma(\Sigma x_i, (n+1)\lambda)$

In particular, if $\Sigma x_i = 83$, $n = 10$, then

$\pi(\lambda|\bar{x}) \sim \Gamma(83, 11)$.

Since $E[\Gamma(\alpha,\lambda)] = \alpha/\lambda$, then $\hat{\lambda} = \frac{83}{11} = 7.64$.

(Bayes estim.)

- For the MLE estimate:

$$\ln L(\lambda|\bar{x}) = 83 \cdot \ln\lambda - 10\lambda$$

$$\frac{d \ln L(\lambda|\bar{x})}{d\lambda} = \frac{83}{\lambda} - 10 = 0 \longrightarrow \hat{\lambda} = \frac{83}{10} = 8.3$$

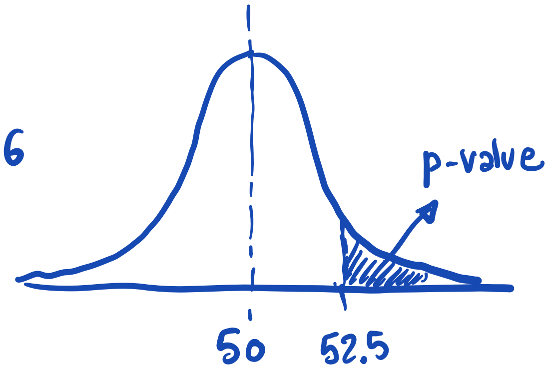
Es 83

(a) p-value $= P[\text{err I type} \mid H_0 \text{ true}]$

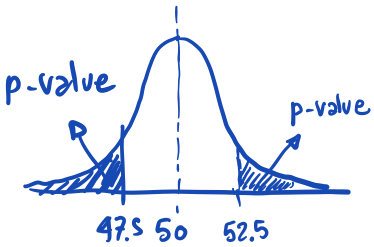
$= P[X_{64} > 52.5 \mid \mu = 50]$

$= P\left[Z > \dfrac{52.5-50}{20/\sqrt{64}}\right] = P[Z>1] = 0.16$

for a 1-sided test.



Otherwise (for 2-sided test)



$= P[\,|X_{64} - 50| > 2.5 \mid \mu = 50]$

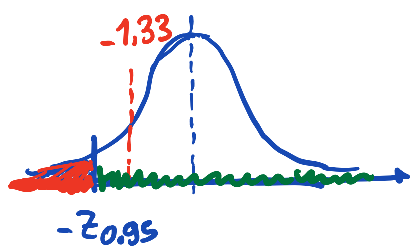
$= P[\,|Z| > 1] = 0.31$

Es 84

(a) $X \sim N(\mu, 1.44)$ $\quad H_0: \mu \geq 7.6 \quad H_1: \mu < 7.6$

$\hat{z}_{obs} = \dfrac{7.2-7.6}{1.2/\sqrt{16}} = -4/3 = -1.33$

$-z_{0.95} = -1.65$



not reject

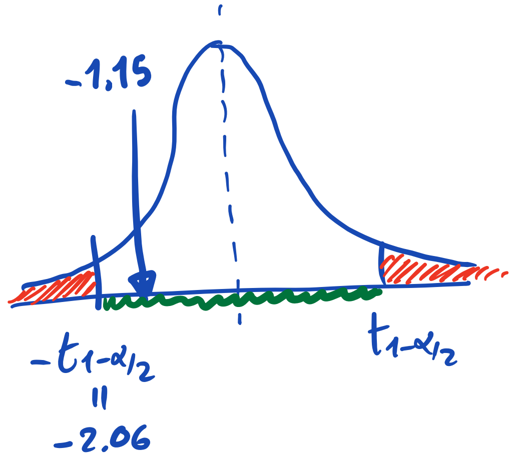
(c) p-value $= \phi(-1.33) = 1 - \phi(1.33) = 1 - 0.908 = 0.092$

Es 85

$n=25$ $\quad H_0: \mu = 20 \quad H_1: \mu \neq 20$

$\bar{X}_n = 19.7 \quad \hat{\sigma}_n^2 = (1.3)^2$

$$\hat{T}_{obs} = \frac{19.7 - 20}{1.3/\sqrt{25}} = -1.15$$



$t_{24,\,0.975} = 2.06$

We can not reject H_0

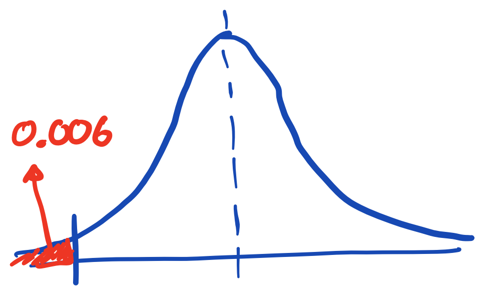
Es 86

$n=10 \quad H_0: \mu \geq 30 \quad H_1: \mu < 30$

$\bar{X}_n = 26.4 \quad \hat{\sigma}_n^2 = 12.3$

$$\hat{T} = \frac{26.4 - 30}{\sqrt{\frac{12.3}{10}}} = -3.24$$

p-value $\sim 1 - 0.994$



Thus .....

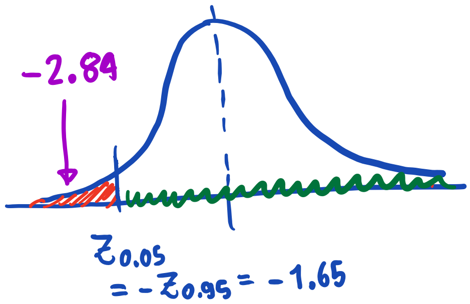
Es 88

Let $H_0: p = 0.45$

(a) $\hat{p} = \frac{70}{200} = 0.35$

If H_0 true : $\bar{X}_n = N\left(0.45, \frac{0.45 \cdot 0.55}{200}\right)$

$$\hat{z} = \frac{0.35 - 0.45}{\sqrt{\frac{0.45 \cdot 0.55}{200}}} = -2.84$$



We reject $H_0: p = 0.45$

Of course we also reject any $p > 0.45$

Es 89

Class	f_i	p_i	$n \cdot p_i$
1 0-30	41	0.45	45
2 30-60	31	0.25	25
3 60-90	13	0.14	13
4 90-∞	15	0.16	15

$p_1 = 1 - e^{-30/50} = 0.45$

$p_2 = e^{-30/50} - e^{-60/50} = 0.25$

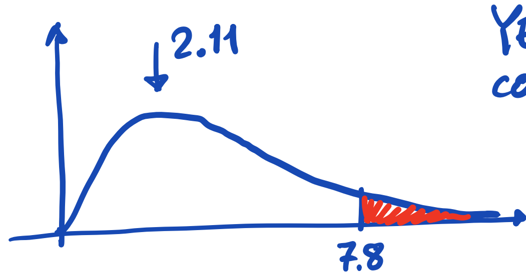
$p_3 = e^{-60/50} - e^{-90/50} = 0.14$

$p_4 = e^{-90/50} = 0.16$

$$W = \sum_{i=1}^{4} \frac{(f_i - np_i)^2}{np_i} = 2.11$$

Let $\alpha = 0.05$

$\chi^2_{4-1, 0.95} = 7.8$



YES, CONSISTENT

Es 90

Y \ X	Alive	Dead
Less	4597	618
Greater	67093	422

Y \ X	Alive	Dead	
Less	0.063	0.0085	0.0717
Greater	0.9225	0.0058	0.9283
	0.9857	0.0143	

$$W = \sum_i \sum_j \frac{(f_{ij} - p_i q_j \cdot n)^2}{p_i q_j \cdot n} = 4327$$

→ too big !!

→ reject independence!

Exercises Inferential Statistics

EXERCISES ON POINT ESTIMATION

(1) Determine the sample size of a random sample from a normal population in such a way that the difference (in absolute value) between the sample mean and the population mean will be less than the 5% of the standard deviation of the population, with a probability greater or equal to 95%. Answer to the same question if the distribution of the population in unknown. (Hint: use Cebicev inequality).

$\big($Answer: $n \geq 1537$; $n \geq 8000\big)$

(2) Find the maximum likelihood estimator of a parameter θ of a population uniformly distributed in $[-\theta, \theta]$ when $X_1, X_2, \ldots, X_n$ has been sampled. Find the estimate if the sampled $3, 5.6, 3, 0, -6, 9, -3.4$ has been observed.

$\big($Answer: $d_{ML}(X_1, \ldots, X_n) = \max_i |X_i|$, $\hat{\theta}_{ML} = 9.$ $\big)$

(3) Compute the maximum likelihood estimator of the unknown parameter θ of a population uniformly distributed in $[0, \theta]$ when $X_1, X_2, \ldots, X_n$ has been sampled and consider the estimator $d_1(X_1, \ldots, X_n) = 2\overline{X}$.

(a) Compare the two estimators mean square errors. Which estimator is preferable?

(b) Find the two estimates if the sample

$$0.31,\ 0.61,\ 0.33,\ 0.21,\ 0.69,\ 0.90,\ 0.71,\ 0.89,\ 0.91,\ 0.11$$

has been observed.

(c) Is the maximum likelihood estimator unbiased? if not, find a linear transformation of it which is unbiased.

(d) Is d_1 unbiased? if not, find the unbiased one.

$\big($Answer: a) d_{ML} is preferable to d_1; b) $\hat{\theta}_{ML} = 0.91$ and $\hat{\theta}_1 = 1.134$; c) d_{ML} is biased, $d = \frac{n+1}{n} d_{ML}$ is unbiased; d_1 is unbiased.$\big)$

(4) Consider a sample $X_1, X_2, \ldots, X_n$ from a population uniformly distributed on $[-\theta, \theta]$ with an unknown parameter $\theta > 0$.

(a) Find the moment estimator for the parameter θ (which is the one obtained by equalizing the sample moments with the corresponding moments in the population). (Hint: use the second moments).

$\big($Answer: a) $d_m(X_1, \ldots, X_n) = \sqrt{3}\sqrt{\frac{\sum_{i:1}^{n} X_i^2}{n}}$$\big)$

(5) $X_1, X_2, \ldots, X_n$ are sampled from a population with p.d.f.

$$f(x; \theta) = \left(\frac{3}{2} + 4\theta\right) x^{4\theta + \frac{1}{2}} \quad \text{if} \quad 0 < x < 1$$

and null otherwise.

(a) Find the maximum likelihood estimator of the parameter θ, ($\theta > -\frac{3}{8}$);

(b) Find the moment estimator for the parameter θ (obtained by equalizing the sample mean to the population mean).

(Answer: a) $d_{ML}(X_1,\ldots,X_n) = -\frac{1}{4}\left(\frac{3}{2} + \frac{n}{\sum\limits_{i:1}^{n}\log X_i}\right)$; b) $d_m(X_1,\ldots,X_n) = \frac{5\overline{X}-3}{8(1-\overline{X})}$)

(6) $X_1, X_2, \ldots, X_n$ are sampled from a population with p.d.f.

$$f(x;\theta) = \frac{1}{2}\theta^3 x^2 e^{-\theta x} \quad \text{if} \quad x > 0$$

and null otherwise.

(a) Find the maximum likelihood estimator of the parameter θ, ($\theta > -\frac{3}{8}$).

(Answer: a) $d_{ML}(X_1,\ldots,X_n) = 3\frac{1}{\overline{X}}$)

(7) Let $X_1, X_2, \ldots, X_n$ be a sample from a population with p.d.f. $f(x;\theta) = \frac{1}{\sqrt{\theta}}e^{-\frac{x}{\sqrt{\theta}}}$ if $x > 0$ and null otherwise.

(a) Find the maximum likelihood estimator of the parameter θ, ($\theta > 0$);

(b) Find the moment estimator for the parameter θ (obtained by equalizing the sample mean to the population mean).

(c) Are the two estimators unbiased? If not, can you find an unbiased estimator?

(Answer: a) $d_{ML}(X_1,\ldots,X_n) = (\overline{X})^2$; b) $d_m(X_1,\ldots,X_n) = (\overline{X})^2$; no, they are biased. $d(X_1,\ldots,X_n) = \frac{n}{n+1}(\overline{X})^2$ is an unbiased estimator of θ.)

(8) Let $X_1, X_2, \ldots, X_n$ be a sample of size n from a population with mean μ and variance $\sigma^2 = 2$. Consider the following estimators of the parameter μ:

$$d_1(X_1,\ldots,X_n) = \overline{X}, \quad d_2(X_1,\ldots,X_n) = \frac{X_1 + \sum\limits_{j:2}^{n-1}(-1)^j X_j + X_n}{2 + \sum\limits_{j:2}^{n-1}(-1)^j}$$

$$d_3(X_1,\ldots,X_n) = \sum_{j:1}^{n}(-1)^j X_j, \quad d_4(X_1,\ldots,X_n) = \frac{2}{n(n+1)}\sum_{j:1}^{n} jX_j$$

(a) Which of the d_k, $k:1,2,3,4$ is an unbiased estimator of μ?

(b) Which of the estimators d_1 and d_4 is preferable?

(Answer: a) d_1, d_2 and d_4 are unbiased estimators of μ, d_3 is not; b) d_1 is preferable to d_4 $n > 1$.)

(9) Let X_1, X_2, X_3, X_4 be a sample from a population with an exponential distribution of parameter $1/\theta$,

$$f(x;\theta) = \begin{cases} \frac{1}{\theta}\mathrm{e}^{-\frac{1}{\theta}x} & x \geq 0 \\ 0 & \text{otherwise} \end{cases}$$

Consider the following estimators of θ

$$d_1 = \frac{1}{2}(X_1 + X_2) + \frac{1}{3}X_3 + \frac{1}{4}X_4, \qquad d_2 = \frac{1}{3}(X_1 + X_2 + X_3) + \frac{1}{4}X_4.$$

(a) Compute the bias of d_1 and the bias d_2.

(b) Compute the mean square errors of the two estimators.

(Answer: a) $\mathbb{E}_\theta(d_1) - \theta = \frac{7}{12}\theta$ and $\mathbb{E}_\theta(d_2) - \theta = \frac{\theta}{4}$; b) $\mathbb{E}_\theta(d_1 - \theta)^2 = \frac{146}{144}\theta^2$ and $\mathbb{E}_\theta(d_2 - \theta)^2 = \frac{11}{24}\theta^2$, d_2 is preferable to d_1.)

(10) Let $X_1, ..., X_n$ be a random sample from a p.m.f.

$$p(x;\theta) = \begin{cases} \frac{\theta^x e^{-\theta}}{x!} & x = 0, 1, 2, ... \\ 0 & \text{otherwise.} \end{cases}$$

(a) Are $d_1 = \sum_{i=1}^n X_i$ and $d_2 = nX_1$ biased estimators of θ? If this is the case, compute the bias.
(b) Which of the two estimators is preferable for estimating θ?

(Answer: a) $\mathbb{E}_\theta(d_i) = n\theta$, and $\mathbb{E}_\theta(d_i) - \theta = (n-1)\theta$, for $i = 1, 2$; b) $\mathbb{E}_\theta(d_1 - \theta)^2 = n\theta + (n-1)^2\theta^2$ and $\mathbb{E}_\theta(d_2 - \theta)^2 = n^2\theta + (n-1)^2\theta^2$, d_1 is preferable.)

(11) $X_1, ..., X_n$ is a random sample from a uniform distribution with unknown parameter θ

$$f(x;\theta) = \begin{cases} \frac{1}{\theta} & 0 \leq x \leq \theta \\ 0 & \text{otherwise} \end{cases}$$

with $\theta > 0$. Consider the estimator $d = \frac{4}{2n-1}\sum_{i=1}^n X_i$ of the parameter θ.
(a) Prove that d is a biased estimator of θ and find an unbiased estimator d'.
(b) Prove that d is asymptotically correct.

(Answer: a) $\mathbb{E}_\theta(d) = \frac{2n}{(2n-1)}\theta \neq \theta$, $d' = \frac{2n-1}{2n}d$; b) $\lim_{n\to\infty} \mathbb{E}_\theta(d) = \theta$.)

(12) Let X_1, X_2 be a random sample from a population with p.d.f.

$$f(x;\theta) = \begin{cases} 2\theta(1+x)^{-(2\theta+1)} & x \geq 0 \\ 0 & \text{otherwise.} \end{cases}$$

(a) Write the likelihood function.
(b) Given that the parameter $\theta \in \{1, 2, 3\}$, find the maximum likelihood estimate of θ having observed the sample $x_1 = 0.1$ of $x_2 = 0.9$.

(Answer: a) $L(\theta; x_1, x_2) = 4\theta^2\{(1+x_1)(1+x_2)\}^{-(2\theta+1)}$; b) $\hat{\theta}_{ML} = 1$.)

EXERCISES ON INTERVAL ESTIMATION

(1) A measuring instrument is used to determine the magnitude of a certain unknown quantity μ. The result of the measurement is the value $X = \mu + \epsilon$ where $\epsilon \sim \mathrm{N}(0, \sigma^2)$. 7 independent measurements are performed and the observed values are: 2.1, 2.2, 1.9, 1.8, 2.3, 2.2, 1.7.
(a) Find the 95 percent two-sided interval for the unknown quantity μ.
(b) Find the 95 percent two-sided interval for σ^2.

(Answer: a) (1.82,2.24); b) (0.022, 0.254))

(2) A random sample of size 16 from a normal population has been observed. If the sample mean and sample variance are respectively 7.9 and 10.433,
(a) find the 95 percent two-sided interval for the mean μ of the population;
(b) find the confidence level to obtain an interval of length 2.5 with these data;
(c) determine the appropriate sample size to obtain a 98 percent confidence interval for μ that is no greater than 2.5 (using the previous data).

(Answer: a) (6.18,9.62); b) $1-\alpha \simeq 85\%$; c) $n \geq 46$)

(3) In a recent study on Internet usage, 248 of the 540 Internet users sampled receive more than 15 emails daily. Compute the proportion in the population of the Internet users who receive daily more than 15 emails and find an approximate 98 percent confidence interval estimate for it.

(Answer: (0.4093, 0.5092))

(4) In August 2013, the New York Times reported that a recent poll indicated that the 52% of the population was in favor of the job performance of President Obama, with a margin of error of ± 4 percent with 95 % confidence. Can we infer how many people were questioned?

(Answer: $n = 600$)

(5) In a packaging company two different machines, A and B, are used. We want to compare the packaging times of the two machines. 10 full packaging processes are observed and the following data are observed:

$$\begin{array}{lllll} A & n=10 & \overline{x}_A = 55 & s_A = 1.4 & s_A^2 = 1.96 \\ B & n=10 & \overline{x}_B = 53 & s_B = 1.5 & s_B^2 = 2.25 \end{array}$$

Assuming that the packaging times in the two populations are normal and have the same variance, determine the 95 percent confidence interval estimate for $\mu_A - \mu_B$.
With the same level of confidence can we state the two samples are observed from the same population?

(Answer: $(0, 637, 3.363)$; no)

(6) It has been estimated that the 36% of the girls which live in big metropolitan city and the 29% who live near the mountains are vegetarian. In both cases, the estimate is based on a sample of 200 girls. Determine a 95 percent confidence interval for the difference of the proportions of vegetarians of the two populations. Can we state that with this confidence level there is no difference?

(Answer: $(-0, 0215, 0.1615)$; yes)

(7) The Ph of a chemical solution is unknown. A set of four measurements yielded the following values of the Ph

$$8.24, \quad 8.18, \quad 8.15, \quad 8.23.$$

Suppose the observations to be $X \sim N(\mu, \sigma^2)$, where μ is the unknown true value of the Ph.

(a) If it is known $\sigma^2 = 0.0025$, find the two-sided 95 percent confidence interval for μ.
(b) What should be the sample size n, if we would like to reduce the length of the interval of $\frac{2}{3}$ of the actual length for the same level of confidence?
(c) Suppose, from now on, σ^2 to be unknown and find the two-sided 95% confidence interval for μ.
(d) Consider, at the same confidence level, the sample with the two additional measurements 7.5 and 8.5 (and sample size $n = 6$), what is the new confidence interval?

(Answer: a) [$8.151,\ 8.249$]; b) 36; c) [$8.1325,\ 8.2674$]; d) [$7.78,\ 8.485$])

(8) From a recent study in Sweden it has been observed that in a sample of 600 residents, 24 are strangers. Find the 95% confidence interval for the proportion p of strangers among the residents in Sweden.

$\big($Answer: $[0.02432,\ 0.05568]$ $\big)$

(9) Suppose the price (in Euro) of a share to be distributed like a normal with mean μ and variance σ^2. In 5 different days it has been observed that the following data $x_1, ..., x_5$

$$\sum_{i=1}^{5} x_i = 32.5, \qquad \sum_{i=1}^{5} x_i^2 = 223.$$

and a confidence interval has been provided for the mean price μ

$$[3.628,\ 9.372].$$

Can you find the level of confidence of the interval?

$\big($Answer: 98% $\big)$

(10) A certain manufacturer produces computer chips. Each chip is independently acceptable with some probability p. To obtain an approximate 99 percent confidence interval for p, whose length is approximately .05, an initial sample of 30 chips has been taken. If 26 of these chips are of acceptable quality, how many *additional* chips should be sampled to get a 99 percent confidence interval approximately of the desired length?

$\big($Answer: 1201 $\big)$

(11) The following are scores on IQ tests of a random sample of 18 students at a large eastern university.

$$130, 122, 119, 142, 136, 127, 120, 152, 141, 132, 127, 118, 150, 141, 133, 137, 129, 142$$

(a) Construct a 95 percent confidence interval estimate of the average IQ score of all students at the university.
(b) Construct a 95 percent lower confidence interval estimate.
(c) Construct a 95 percent upper confidence interval estimate.

$\big($Answer: a) $(128.12; 138.28)$; b) $(0, 137.38)$; $(129.01, \infty)\big)$

(12) A market research firm is interested in determining the proportion of households that are watching a particular sporting event. To accomplish this task, they plan on using a telephone poll of randomly chosen households. How large a sample is needed if they want to be 90 percent certain that their estimate is correct to within ± 0.2?

$\big($Answer: $n \geq 17$ $\big)$

(13) To estimate p, the proportion of all newborn babies that are male, the gender of 10000 newborn babies was noted. If 5106 of them were male, determine:
(a) a 90 percent confidence (approximate) interval estimate for p;
(b) a 99 percent confidence (approximate) interval estimate for p.

$\big($Answer: a) $(0.502, 0.519)$; b) $(0.498, 0.523)$ $\big)$

(14) (*) Consider two independent samples from two normal populations having the same variance σ^2, respectively of size n and m. Denote by S_1^2 and S_2^2 the respective sample variances. Define the *pooled estimator*

$$S_p^2 = \frac{(n-1)S_1^2 + (m-1)S_2^2}{n+m-2}.$$

(a) Is S_p^2 an unbiased estimator of σ^2?
(b) Show that S_p^2 is the minimum mean square estimator of σ^2 of the form

$$d_\lambda = \lambda S_1^2 + (1-\lambda)S_2^2, \quad \lambda \in (0,1).$$

(Recall $\mathrm{Var}(\chi_k^2) = 2k$).

$\big($Answer: a) yes; b) d_λ are unbiased, $\mathrm{Var}(d_\lambda)$ is minimum at $\lambda = \frac{n-1}{n+m-2}$.$\big)$

(15) A sample of 9 measures observed from a normally distributed population assumed the following values:

$$0.2;\ 1.4;\ 2.3;\ 0.6;\ 2.5;\ -1.3;\ 0.8;\ -1.8;\ -0.2.$$

(a) Provide an interval estimate of the mean, with a confidence level $\alpha = 0.1$, assuming that the variance is known and equal to 4.
(b) Provide an interval estimate of the mean, with a confidence level $\alpha = 0.1$, assuming that the variance is unknown.
(c) Provide an interval estimate of the variance with a confidence level $\alpha = 0.1$.

$\big($Answer:$\big)$ $\overline{x} = 0.5, \ \ s^2 = 2.16$

a) $z_{0.95} = 1.65$, $I = [0.5 - 1.65\frac{\sqrt{4}}{\sqrt{9}}, 0.5 + 1.65\frac{\sqrt{4}}{\sqrt{9}}] = [-0.60, 1.60]$.
b) $t_{0.95} = 1.86$ (8 df), $I = [0.5 - 1.86\frac{\sqrt{2.16}}{\sqrt{9}}, 0.5 + 1.86\frac{\sqrt{2.16}}{\sqrt{9}}] = [-0.41, 1.41]$.
c) $q_{0.95} = 15.51$, $q_{0.05} = 2.73$ (8 df), $I = [\frac{8\cdot 2.16}{15.51}, \frac{8\cdot 2.16}{2.73}] = [1.11, 6.32]$.

EXERCISES ON HYPOTHESIS TESTING

(1) A random sample $X_1, \ldots, X_{20}$ from a normal population with mean unknown μ and variance σ^2) has been observed and

$$\sum_{i=1}^{20} x_i = 65 \qquad \sum_{i=1}^{20} x_i^2 = 530.$$

(a) given $\sigma^2 = 16$, would you reject with $\alpha = 0.05$ the null hypothesis

$$H_0 : \ \mu = 2 \quad \text{vs} \quad H_1 : \ \mu = 4?$$

(b) answer to the same question if the value of σ^2 is unknown.
(c) Given the critical region $C = \{\bar{X}_{20} \geq 3.2175\}$, find the significance level associated to the test if σ^2 is unknown.

$\big($Answer: a) no, H_0 is accepted; b) H_0 is accepted; c) $\alpha = 0.1\big)$

(2) A random sample $X_1, \ldots, X_{10}$ from a normal population is observed and

$$\sum_{i=1}^{10} x_i = 4.91 \qquad \sum_{i=1}^{10} x_i^2 = 48.$$

Find the p-value associated to the sample for the hypotheses

$$H_0 : \mu = 2 \quad \text{vs} \quad H_1 : \mu < 2.$$

(Answer: p-value= 0.025)

(3) Given the random sample $(-4.4, 4.0, 2.0, -4.8)$ sampled from a normal population with mean μ and variance σ^2,

(a) determine an interval with confidence level 99% for the mean μ;

(b) Given the sample observed, would you reject or accept

$$H_0 : \mu = -1 \quad \text{vs} \quad H_1 : \mu \neq -1$$

at a significance level $\alpha = 0.01$?

(Answer: a) $[-13.84, 12.24]$; b) H_0 is accepted)

(4) In a normal population with variance equal to 4, the null hypothesis $\mu = 10$ is tested against a one-sided alternative with a level of significance equal to 5%.

(a) Find the sample size in order to have the probability of the second type error less or equal to 5% if the alternative hypothesis $\mu = 7$ is true.

(Answer: a) $n \geq 5$)

(5) In testing the null hypothesis $\mu = 11$ versus $\mu > 11$ with a probability of the type I error equal to 5%, it has been found an estimate of the variance of the population equal to 4 from a sample of 20 observations. Find the sample size that would be necessary to have a probability of the type II error equal to 10% if the alternative hypothesis $\mu = 12$ is true.

(Answer: $n \geq 38$)

(6) An airline company created a new customer-loyalty program, believing that 5% of its customers would have enrolled in the program. From a random sample of 500 customers there have been 34 enrollments.

(a) Using these data test the null hypothesis that the company hypothesis is correct against the alternative that it is not with a significance level of 0.01.

(b) The test has been used, what is the probability that the null hypothesis has been accepted given the true percentage of customers enrolled in the program was 10%?

(Answer: a) do not reject H_0; b) $\beta(0.1) = 0.03$)

(7) Given a sample of size 100 with $\overline{x} = 2.7$ and $\sum_{i:1}^{100}(x_i - \overline{x})^2 = 225$, test the hypotheses $H_0 : \mu = 3$ and $H_0 : \sigma^2 = 2.5$ with a level of significance of 0.01 (with a suitable alternative hypothesis and assuming the normality of the population). What is the probability of mistakenly accepting $H_0 : \mu = 3$ when instead the true value is $\mu = 2.5$?

(Answer: in both cases I do not reject H_0 ($t_{\text{comp}} = -1.99, \chi^2_{\text{comp},99} = 90$; $\beta(2.5) = 23\%$)

(8) Two insect sprays (spray 1 and spray 2) are compared. The percentage of insects killed, on a given surface and in a fixed time interval, with the spray 1 has been equal to 0.64 with a sample size of 350 individuals, whereas the percentage has been equal to 0.52 with a sample size of 300 insects with the spray 2.
(a) With a significance level $\alpha = 0.05$, can you reject the hypothesis that the two sprays are equivalent?
(b) Compute the p-value.

$\big($Answer: yes; p-value$= 0.2\%$ $\big)$

(9) Two blood pressure meters (A and B) gave the following measurements on a sample of 16 individuals:

TYPE A	79, 75, 93, 84, 93, 56, 87, 86,
TYPE B	66, 74, 82, 77, 91, 58, 79, 80

(a) Is there a meaningful difference between the mean of measurements of the two devices? (assume $\alpha = 0.05$)
(b) Compute the probability of the error of type II for the alternative hypotheses $\mu_A - \mu_B = 3$, $\mu_A - \mu_B = -3$

$\big($Answer: a) no; b) $\beta(3) = .92$, $\beta(-3) = .92\big)$

(10) For solving the test

$$H_0 : \ p = 0.4 \quad \text{vs} \quad H_1 : \ p = 0.7$$

on the parameter p of a Bernoulli population with a sample of size 3, the critical region

$$C := \{\sum_{i=1}^{3} X_i \geq 2\}.$$

has been proposed.
(a) Given the sample $(x_1, x_2, x_3) = (1, 1, 0)$, will H_0 be accepted?
(b) Compute the probability of the error of type I.
(c) Compute the power function of the test and its value for $p = 0.7$.

$\big($Answer: a) no; b) $\alpha \approx 0.352$; c) $1 - \beta(p) = 3p^2 - 2p^3$, $1 - \beta(0.7) = 0.784\big)$

(11) To check if the Internet service provider AT&T has increased its market share from 15 percent to 20 percent in a certain area, a random sample of 1000 Internet customers in this area are polled and it yielded that 270 are AT&T customers. Would you conclude at the 5 percent level of significance that the Internet provider has increased its market share?

$\big($Answer: $H_0 : p = 0.15$, $H_1 : p = 0.2$, H_0 is rejected$\big)$

(12) Consider the two hypotheses

$$H_0 : \theta = 2 \quad \text{vs} \quad H_1 : \theta = 4$$

for the parameter of a Poisson. Given one observation X_1 the critical region $C = \{X_1 : X_1 = 3\}$ has been proposed.

Compute the probabilities of the error of type I and type II.

$\big($Answer: $\alpha \approx 0.18$, $\beta \approx 0.80\big)$

(13) The Green association has decided to analyze the level of mercury (mg/kg) in the fishes of a lake. 10 fishes has been sampled and the following values have been observed

$$0.8,\ 1.6,\ 0.9,\ 0.8,\ 1.2,$$
$$0.4,\ 0.7,\ 1.0,\ 1.2,\ 1.1.$$

Suppose the quantity of mercury in a fish to be a random variable normally distributed with mean μ and variance σ^2.

(a) Would you reject with a significance level $\alpha = 0.05$ the hypothesis

$$H_0 : \ \mu = 0.9 \quad \text{vs} \quad H_1 : \ \mu = 1.0?$$

(b) If the critical region

$$C = \{\bar{X}_{10} \geq 0.9 + (1.02774)S\},$$

has been proposed, what is the significance level associated to the test?

$\big($Answer: H_0 accepted; $\alpha = 0.005\big)$

(14) Given the sample

$$0.4;\ \ 1.2;\ \ -0.5;\ \ 0.8;\ \ 2.0;\ \ -1.1;$$

extracted from a normally distributed population, test the hypothesis $\mu = 0$ versus $\mu \neq 0$

a) with a confidence level $\alpha = 0.05$ assuming the variance known and equal to 1;

b) with a confidence level $\alpha = 0.05$ assuming the variance unknown.

c) Test the hypothesis that the variance is equal 2 with a confidence level $\alpha = 0.05$.

$\big($Answer:$\big)$ $\overline{x} = 0.47, \ \ s^2 = 1.28$

a) $Z = \frac{0.47-0}{1/\sqrt{6}} = 1.16$, $z_{0.975} = 1.96$, $C' = (-\infty, -1.96) \cup (1.96, \infty)$. $\longrightarrow$ Accept H_0.

b) $T = \frac{0.47-0}{\sqrt{1.28}/\sqrt{6}} = 1.01$, $t_{0.975} = 2.57$ (5 df), $C' = (-\infty, -2.57) \cup (2.57, \infty)$. $\rightarrow$ Accept H_0.

c) $Q = \frac{5 \cdot 1.28}{2} = 3.2$, $q_{0.025} = 0.831$, $q_{0.975} = 12.83$ (5 df), $C' = [0, 0.831) \cup (12.83, \infty)$. $\longrightarrow$ Accept H_0.

(15) A sample of 10 measures assumed the following values:

$$1.2;\ 0.4;\ 2.3;\ 0.6;\ 2.5;\ -1.3;\ 0.8;\ -2.8;\ -1.2; -0,8.$$

(a) Check the normality of the population by means of a Kolmogorov-Smirnov test.

(b) Test the hypothesis $\mu = 0$ with a confidence level $\alpha = 0.1$, assuming that the standard deviation is known and equal to 1.

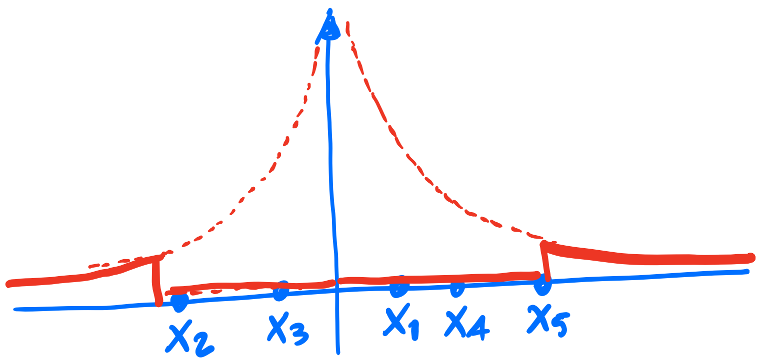
(c) Test the hypothesis $\sigma^2 = 2$ with a confidence level $\alpha = 0.1$.

(16) To test the regularity of a dice we toss it 100 times. Doing it, we counted the following occurrences of faces of the dice: 21 (for 1), 16 (for 2), 15 (for 3), 13 (for 4), 22 (for 5), 13 (for 6). Test, with an appropriate tool, the hypothesis that the dice is not loaded (i.e., it is regular).

SOME SOLUTIONS

EX2 - pointwise estimates

$$L(\theta|\bar{x}) = \prod_{i=1}^{n} \frac{1}{2\theta} \cdot \mathbb{1}_{[-\theta,\theta]}(x_i) = (2\theta)^{-n} \cdot \prod_{i=1}^{n} \mathbb{1}_{[-\theta,\theta]}(x_i)$$



graph of $L(\theta|\bar{x})$. Observe it is $=0$ in the interval

$$\left[-\max\{|x_i|\}, +\max\{|x_i|\}\right]$$

(because in this interval at least one of the $\mathbb{1}_{[-\theta,\theta]}(x_i)$ is $=0$)

Thus, the max of $L(\theta|\bar{x})$ is reached

in $\hat{\theta} = \max\{|x_i|\} = 9$

(we have seen a similar exercise in class, for MLE of $X \sim f(x) = \frac{2x}{\theta^2} \cdot \mathbb{1}_{[0,\theta]}(x)$)

Ex3 - pointwise estimates

The MLE estimator of θ for $X \sim U[0,\theta]$ is

$\hat{\theta} = \max\{X_i\}$ (see lectures, or exercise 2)

Let $\hat{d}_1 = 2\bar{X}_n$

(a) $F_{\hat{\theta}}(x) = \left(\frac{x}{\theta}\right)^n$ $\quad f_{\hat{\theta}}(x) = \frac{n x^{n-1}}{\theta^n}$ $\quad E[\hat{\theta}] = \int_0^\theta \frac{x \cdot n x^{n-1}}{\theta^n} dx = \frac{n}{\theta^n}\left[\frac{x^{n+1}}{n+1}\right]_0^\theta$

$= \frac{n}{n+1}\theta$

$$V[\hat{\theta}] = E[\hat{\theta}^2] - E[\hat{\theta}]^2$$
$$= \int_0^\theta \frac{x^2 n x^{n-1}}{\theta^n} dx - \left(\frac{n}{n+1}\right)^2 \theta^2 = \frac{n}{n+2}\theta^2 - \frac{n^2}{(n+1)^2}\theta^2 = \theta^2\left[\frac{n}{(n+2)(n+1)^2}\right]$$

$$E[\hat{d}_1] = 2E[\bar{X}_n] = 2 \cdot \frac{\theta}{2} = \theta$$
$$V[\hat{d}_1] = 4 \cdot V[\bar{X}_n] = \frac{4}{n^2}\left[\frac{\theta^2}{12} + \dots + \frac{\theta^2}{12}\right] = \theta^2 \cdot \left[\frac{3}{n^2}\right]$$

$$MSE(\hat{\theta}) = V[\hat{\theta}] + (E[\hat{\theta}] - \theta)^2 = \text{----}$$
$$MSE(\hat{d}_1) = V[\hat{d}_1] + (E[\hat{d}_1] - \theta)^2 = \text{----}$$

(then compare)

(b) $\hat{\theta} = 0.91$ $\qquad \hat{d}_1 = 2 \cdot \left[\frac{\sum X_i}{n}\right] = 0.61$

(c) $\hat{\theta}$ is biased, being $E[\hat{\theta}] \neq \theta$

but $\hat{\theta}_2 = \frac{n+1}{n} \cdot \hat{\theta}$ is unbiased

(d) $\hat{d}_1$ is unbiased, being $E[\hat{d}_1] = \theta$

Ex 8 point estimates (estimator $\hat{d}_4$)

Recall that $\sum_{j=1}^{n} j = \frac{n \cdot (n+1)}{2}$.

It follows

$$E[d_4] = E\left[\frac{2}{n(n+1)} \sum^{n} j \cdot X_i\right] = \frac{2}{n(n+1)} \cdot \sum_{j=1}^{n} j \cdot \mu = \mu$$

$$V[d_4] = V\left[\frac{2}{n(n+1)} \sum^{n} j \cdot X_i\right]$$

$$= \left(\frac{2}{n(n+1)}\right)^2 \cdot \sum^{n} V[j \cdot X_i]$$

$$= \left(\frac{2}{n(n+1)}\right)^2 \cdot \sum^{n} j^2 \cdot 2 = \frac{8}{n(n+1)} \cdot \frac{(2n+1)}{6}$$

(from $\sum_{j=1}^{n} j^2 = \frac{n \cdot (n+1)(2n+1)}{6}$)

Ex 10 point estimates (point b)

$$E[(\hat{d}_1-\theta)^2] = Var(\hat{d}_1) + (\text{bias}\quad)^2$$
$$= Var(\hat{d}_1) + (n-1)^2\theta^2 = n\theta + (n-1)^2\theta^2$$

[Note that $\hat{d}_1 \sim Poiss(n\theta)$ (is sum of indep. Poisson)
thus $Var(\hat{d}_1) = n\theta$]

$$E[(\hat{d}_2-\theta)^2] = Var(\hat{d}_2) + (\text{bias})^2$$
$$= n^2\theta + (n-1)^2\theta^2$$

[$Var(\hat{d}_2) = Var(n \cdot X) = n^2 \cdot Var(X) = n^2 \cdot \theta$
being $X \sim Poiss(\theta)$]

Ex 2 Interval estimation

$X \sim N \qquad n = 16 \qquad \bar{X}_n = 7.9 \qquad \hat{\sigma}_u^2 = 10433$

a) $\alpha = 0.05 \to 1 - \alpha/2 = 0.975$

small sample, from normal, $\to t_{n-1,1-\alpha/2} = t_{15,0.975} = 2.13$

$$I_{1-\alpha} = \left[\bar{X}_n - t_{n-1,1-\alpha/2}\sqrt{\frac{\hat{\sigma}^2}{n}}, \bar{X}_n + t_{n-1,1-\alpha/2}\sqrt{\frac{\hat{\sigma}^2}{n}}\right]$$

$$= \left[7.9 - 2.13\sqrt{\frac{10433}{16}}, 7.9 + 2.13\cdot\sqrt{\frac{10433}{16}}\right] = [6.18, 9.68]$$

b) range $I_{1-\alpha} = 2.5 \Rightarrow 2\, t_{n-1,1-\alpha/2}\cdot\sqrt{\frac{\hat{\sigma}^2}{n}} = 2.5$

$\Rightarrow t_{15,1-\alpha/2} = \frac{2.5}{2}\cdot\frac{4}{3.23} = 1.54$

$\Rightarrow$ (table) $1-\alpha/2 \sim 0.925$

$\Rightarrow 1-\alpha = 0.85$

c) $1-\alpha = 0.98 \Rightarrow 1-\alpha/2 = 0.99$

range $I_{0.98} \leq 2.5 \Rightarrow 2\, t_{15,0.99}\cdot\sqrt{\frac{\hat{\sigma}_n^2}{n}} \leq 25$

$\Rightarrow t_{15,0.99} \sim 2.6$

(or approx with $z_{0.99} = 2.32$)
(if "n" is large)

$$\frac{2\cdot 2.6\cdot\sqrt{10433}}{2.5} \leq \sqrt{n}$$

$\longrightarrow n \geq 45.13$

(little different if one uses z)

Ex 4 Interval estimates

$X_i = \begin{cases} 1 & p \quad (\text{favorable}) \\ 0 & 1-p \quad (\text{contrary}) \end{cases}$ $\quad X_i \sim Bern(p) \qquad \hat{p} = 0.52 \qquad \hat{p}(1-\hat{p}) \simeq 0.25$

$$z_{1-\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = 0.04 \quad \text{with} \quad \alpha = 0.05 \rightarrow 1-\alpha/2 = 0.975$$

$$\Rightarrow \underbrace{z_{0.975}}_{(1.96)}\sqrt{\frac{0.25}{n}} = 0.04 \quad \Rightarrow \quad \frac{0.25}{n} = \left(\frac{0.04}{1.96}\right)^2 \quad \Rightarrow \quad n \sim 600$$

Ex 10 Interval estimates

$X_i = \begin{cases} 1 & p \quad (\text{acceptable}) \\ 0 & 1-p \end{cases}$ $\quad X_i \sim Bern(p), \quad \hat{p} = \frac{26}{30} \sim 0.867, \quad \hat{p}(1-\hat{p}) \simeq 0.12$

range interval $= 2 \cdot z_{1-\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$ where $n = 30 + n_1$

where $n_1 =$ n° chip added

$$\alpha = 1 - 0.99 = 0.01 \quad \Rightarrow \quad 1-\alpha/2 = 0.995$$

If range $= 0.05$, then

$$2 \cdot \underbrace{z_{0.995}}_{2.575} \cdot \sqrt{\frac{0.12}{n}} = 0.05$$

$$\Rightarrow \frac{0.12}{n} = \left(\frac{0.05}{2 \cdot 2.575}\right)^2 \quad \Rightarrow \quad n = 0.12 \cdot \left(\frac{0.05}{2 \cdot 2.575}\right)^2 \sim 1.275$$

$$n_1 = 1273 - 30 = 1243$$

Ex 12 Interval estimates

Consider the confidence interval for Bernoulli

$$I = \left[\bar{X}_n - z_{1-\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}},\ \bar{X}_n + z_{1-\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \right]$$

Thus it must be $z_{0.95}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \leq 0.2$

Here one must observe that $\forall \hat{p} \in (0,1)$ it holds $\hat{p}(1-\hat{p}) \leq 1/4$.

Also, $z_{0.95} = 1.65$. Thus

$$1.65 \cdot \sqrt{\frac{1}{4n}} \leq 0.2 \quad \Rightarrow \quad \frac{1.65}{0.2} \leq 2\sqrt{n}$$

$$\Rightarrow \quad \frac{1.65}{0.4} = \sqrt{n}$$

$$\Rightarrow \quad \sqrt{n} = 4.125$$

$$\Rightarrow \quad n = 17.01$$

$$\Rightarrow \quad n \geq 17$$

Ex 14 Interval estimates (DIFFICULT! NOT FOR EXAMS)

(a) $E[S_p^2] = \frac{1}{n+m-2} \cdot E[(n-1)S_1^2 + (m-1)S_2^2]$

One must remember that $\frac{(n-1)S_1^2}{\sigma^2} \sim \chi^2_{n-1}$, so that

$$= \frac{1}{n+m-2} \cdot \sigma^2 \cdot E[\chi^2_{n-1} + \chi^2_{m-1}]$$

$$= \frac{1}{n+m-2} \cdot \sigma^2 E[Z_1^2 + Z_2^2 + \dots + Z_{n+m-2}^2]$$

$$= \frac{1}{n+m-2} \cdot \sigma^2 \cdot (\underbrace{1+1+\dots+1}_{n+m-2 \text{ times}}) = \sigma^2$$

(b) $V[\lambda S_1^2 + (1-\lambda)S_2^2] = \lambda^2 V[S_1^2] + (1-\lambda)^2 V[S_2^2]$

$$= \lambda^2 \cdot \frac{\sigma^4}{(n-1)^2} \cdot V[\chi^2_{n-1}] + (1-\lambda)^2 \cdot \frac{\sigma^4}{(m-1)^2} \cdot V[\chi^2_{m-1}]$$

$$= \sigma^2 \cdot \left[\frac{\lambda^2}{(n-1)^2} \cdot 2(n-1) + \frac{(1-\lambda)^2}{(m-1)^2} \cdot 2(m-1) \right]$$

$$= 2\sigma^2 \left[\frac{\lambda^2}{(n-1)} + \frac{(1-\lambda)^2}{(m-1)} \right] = 2\sigma^2 h_{n,m}(\lambda)$$

One must now look for the minimum of $h_{n,m}(\lambda)$, which is reached in $\operatorname{argmin} h_{n,m}(\lambda) = \frac{n-1}{n+m-2}$

Ex 1 Hypothesis testing

a) The rejection region for unilateral tests

$$H_0: \ \mu = \mu_0 \quad \text{vs} \quad H_1: \ \mu = \mu_1 \qquad \text{with} \qquad \mu_0 < \mu_1,$$

with know variance, is

$$\mathcal{R} = \left\{ (x_1, ..., x_n) : \ \bar{x}_n \geq \mu_0 + z_{1-\alpha} \frac{\sigma}{\sqrt{n}} \right\}.$$

Here $\alpha = 0.05 \quad \Rightarrow \quad z_{1-\alpha} = z_{0.95} = 1.645$

Thus, the rejection region is

$$\mathcal{R} = \left\{ (x_1, ..., x_{20}) : \ \bar{x}_{20} \geq 2 + 1.645 \cdot \frac{4}{\sqrt{20}} \right\}$$
$$= \{ (x_1, ..., x_{20}) : \ \bar{x}_{20} \geq 3.47 \}.$$

Since

$$\bar{x}_{20} = 3.25 < 3.47$$

it follows $(x_1, ..., x_{20}) \notin \mathcal{R}$, thus H_0 is not rejected

b) With known σ^2 we have

$$\mathcal{R} = \left\{ (x_1, ..., x_n) : \ \bar{x}_n \geq \mu_0 + t_{1-\alpha}^{(n-1)} \frac{S}{\sqrt{n}} \right\}.$$

- $S = \sqrt{\frac{1}{20-1} \left(\sum_{i=1}^{20} x_i^2 - 20 \, (\bar{x}_{20})^2 \right)} = 4.1$
- $\alpha = 0.05 \quad \Rightarrow \quad t_{1-\alpha}^{(n-1)} = t_{0.95}^{(19)} = 1.729$

Thus the rejection region is

$$\mathcal{R} = \left\{ (x_1, ..., x_{20}) : \ \bar{x}_{20} \geq 2 + 1.729 \cdot \frac{4.1}{\sqrt{20}} \right\}$$
$$= \{ (x_1, ..., x_{20}) : \ \bar{x}_{20} \geq 3.59 \}.$$

Since

$$\bar{x}_{20} = 3.25 < 3.59$$

it follows $(x_1, ..., x_{20}) \notin \mathcal{R}$, thus H_0 is not rejected

c) To find the value α, given the rejection region

$$\mathcal{R} = \{(x_1, ..., x_{20}) : \ \bar{x}_{20} \geq \underbrace{3.2175}_{\mu_0 + t_{1-\alpha}^{(n-1)} \frac{S}{\sqrt{n}}} \}$$

we have

$$3.2175 = \mu_0 + t_{1-\alpha}^{(n-1)} \frac{S}{\sqrt{n}} = 2 + t_{1-\alpha}^{(19)} \frac{4.1}{\sqrt{20}} = 2 + t_{1-\alpha}^{(19)} 0.9168.$$

It follows

$$t_{1-\alpha}^{(19)} = \frac{3.2175 - 2}{0.9168} = 1.328$$

Since

$$t_{0.90}^{(19)} = 1.328,$$

it follows the value $\alpha = 0.1.$

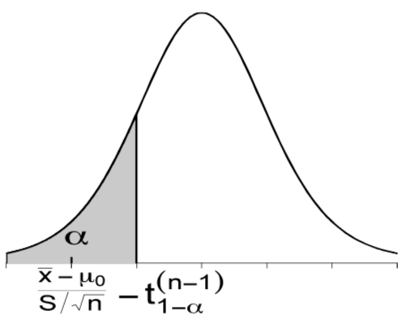
Ex 2 Hypotesis Testing

Small sample, normal population. The rejection region is

$$
\begin{aligned}
\mathcal{R} &= \left\{ (x_1, \dots, x_n) : \bar{x}_n \leq \mu_0 - t^{(n-1)}_{1-\alpha} \frac{S}{\sqrt{n}} \right\} \\
&= \left\{ (x_1, \dots, x_n) : \frac{\bar{x}_n - \mu_0}{S/\sqrt{n}} \leq -t^{(n-1)}_{1-\alpha} \right\}
\end{aligned}
$$

which is based on

$$
P\left(\underbrace{\frac{\bar{X}_n - \mu_0}{S/\sqrt{n}}}_{\sim t^{(n-1)} \text{ under } H_0} \leq -t^{(n-1)}_{1-\alpha} \right) = \alpha \quad \text{under } H_0
$$



move $-t^{(n-1)}_{1-\alpha}$ until $\frac{\bar{x}_n - \mu_0}{S/\sqrt{n}}$ and find the corresponding α

Here we have

$$
S = \sqrt{\frac{1}{9} \left(\sum_{i=1}^{10} x_i^2 - 10 (\bar{x}_{10})^2 \right)} = 2.11
$$

Thus the p-value is

$$
\begin{aligned}
P\left(\underbrace{\frac{\bar{X}_n - \mu_0}{S/\sqrt{n}}}_{t^{(n-1)}} \leq \underbrace{\frac{\bar{x}_n - \mu_0}{S/\sqrt{n}}}_{\text{quantile } t\text{-student}} \right) &= P\left(t^{(9)} \leq \frac{0.491 - 2}{2.11/\sqrt{10}} \right) \\
&= P\left(t^{(9)} \leq -2.262 \right) = 1 - P\left(t^{(9)} \leq \underbrace{2.262}_{t^{(9)}_{1-\alpha}} \right) = 1 - 0.975 = 0.025
\end{aligned}
$$

Ex 4 Hypothesis testing

Since the alternative is $\mu=7$, then the acceptance region will be of the kind $[a, \infty)$

Under H_0: $\bar{X}_n \sim N(10, 4/n)$

$\longrightarrow$ accept if $\bar{X}_n \in [10 - z_{0.95} \cdot \sqrt{4/n}, +\infty)$

$\bar{X}_n \in [10 - 1.65 \cdot 2/\sqrt{n}, +\infty)$

$\bar{X}_n \in [10 - 3.3/\sqrt{n}, +\infty)$

Under H_1: $\bar{X}_n \sim N(7, 4/n)$

$\beta = P[\text{accept } H_0 \mid H_1 \text{ is true}] = P[\bar{X}_n > 10 - 3.3/\sqrt{n} \mid \bar{X}_n \sim N(7, 4/n)] \leq 0.05$

$$P\left[\frac{\bar{X}_n - 7}{\sqrt{4/n}} > \frac{10 - 3.3/\sqrt{n} - 7}{\sqrt{4/n}}\right] = P[Z > z_{0.95}] = 0.05$$

Thus $$\frac{3 - 3.3/\sqrt{n}}{2/\sqrt{n}} = 1.65$$

$$3 - 3.3/\sqrt{n} = 3.3/\sqrt{n}$$

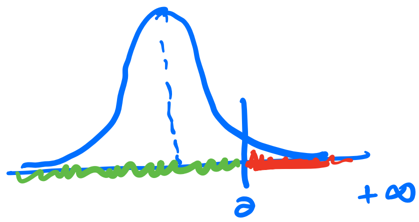
$$3 = 6.6/\sqrt{n} \qquad \sqrt{n} = \frac{6.6}{3} = 2.2$$

$$n = (2.2)^2 = 4.84 \longrightarrow n \geq 5$$

Ex 5 - Hypothesis testing

The test is unilateral ($H_0: \mu = 11$, $H_1: \mu > 11$), thus the rejection for $T_n = \dfrac{\bar{X}_n - 11}{\sqrt{\hat{\sigma}^2/n}}$ is of kind $[a, +\infty)$ (reject H_0 in favor of H_1 if T_n is "too big").

If $\alpha = 0.05$ then $a = t_{n-1, 0.95}$



Assuming n "large", let us approx with Z $\longrightarrow$ $a = z_{0.95} = 1.65$

We want

$$0.10 = P[\text{accept } H_0 \mid H_1 \text{ true}] = P[T_n < 1.65 \mid \mu = 12]$$

$$= P\left[\frac{\bar{X}_n - 11}{\sqrt{4/n}} < 1.65 \;\middle|\; \bar{X}_n \sim N(12, 4)\right]$$

$$= P\left[\frac{\bar{X}_n - 12}{\sqrt{4/n}} + \frac{1}{\sqrt{4/n}} < 1.65\right]$$

$$= P\left[Z < 1.65 - \frac{1}{\sqrt{4/n}}\right] = 0.10$$

$$\Rightarrow \quad 1.65 - \frac{1}{\sqrt{4/n}} = z_{0.10} = -z_{0.90} = -1.28$$

$$\Rightarrow \quad 1.65 + 1.28 = \sqrt{\frac{n}{4}} \quad \Rightarrow \quad \sqrt{n} = 5.86$$

$$\Rightarrow \quad n \geq 34$$

(approx: it would be used t)

Ex 10 Hypothesis testing

a) Since $\sum_{i=1}^{3} x_i = 2$, then the sample (1,1,0) is such that $(1,1,0) \in \mathcal{R} \quad \Rightarrow \quad H_0$ is rejected

b) The probability of an error of I kind is given by

$$\alpha = P\Big((X_1, X_2, X_3) \in \mathcal{R} \mid H_0 \text{ true}\Big)$$

$$= P\Big(\underbrace{\sum_{i=1}^{3} X_i}_{\text{Binom}(3,\theta)} \geq 2 \Big| \theta = 0.4\Big) \qquad \qquad \sum_{i=1}^{3} X_i \in \{0,1,2,3\}$$

$$= P\Big(\sum_{i=1}^{3} X_i = 2 \Big| \theta = 0.4\Big) + P\Big(\sum_{i=1}^{3} X_i = 3 \Big| \theta = 0.4\Big)$$

$$= \binom{3}{2}(0.4)^2(0.6)^1 + \binom{3}{3}(0.4)^3(0.6)^0 \approx 0.352,$$

c) The power function is

$$\pi(\theta) = P(\text{“ reject ” } H_0 | \theta) \qquad \forall \theta$$

$$= P\Big((X_1, X_2, X_3) \in \mathcal{R} \mid \theta\Big)$$

$$= P\Big(\sum_{i=1}^{3} X_i \geq 2 \Big| \theta\Big)$$

$$= \binom{3}{2}\theta^2(1-\theta)^1 + \binom{3}{3}\theta^3(1-\theta)^0 = \theta^3 + 3\theta^2(1-\theta)$$

Ex 11 - Hypothesis testing

We must test

$$H_0: \ p = p_0 = 0.15 \quad \text{vs} \quad H_1: \ p = p_1 = 0.2$$

at a level $\alpha = 0.05$. Let us define

$$X_i = \begin{cases} 1 & i-th \text{ customer with AT\&T} \\ 0 & \text{otherwise} \end{cases}$$

Thus, under H_0, $\rightarrow$ $X_i \overset{iid}{\sim} \text{Bern}(p_0)$

And, for the Central Limit Theorem, if H_0 is true

$$\frac{\bar{X} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} \approx N(0,1) .$$

Thus, the approximate rejection region is

$$\mathcal{R} = \left\{ (x_1, ..., x_{1000}) : \ \bar{x} \geq p_0 + z_{1-\alpha} \sqrt{\frac{p_0(1-p_0)}{n}} \right\}$$

$$= \left\{ (x_1, ..., x_{1000}) : \ \bar{x} \geq 0.15 + z_{1-\alpha} \sqrt{\frac{(0.15) \cdot (0.85)}{1000}} \right\}$$

where $z_{1-\alpha} = z_{0.95} = 1.645$,

$$\mathcal{R} = \{ (x_1, ..., x_{1000}) : \ \bar{x}_{1000} \geq 0.1686 \}.$$

Since

$$\bar{x}_{1000} = 0.27 > 0.1686 \qquad \Rightarrow \qquad (x_1, ..., x_{1000}) \in \mathcal{R},$$

the hypothesis H_0 is rejected.

Ex 12 - Hypothesis testing

In a Hypothesis testing there are 2 kinds of errors

	H_0 true	H_0 false
reject H_0	err I kind	CORRECT
accept H_0	CORRECT	err II kind

The probability of an error of I kind is

$$\alpha = P(X_1 \in \mathcal{R} \mid H_0 \text{ vera})$$

$$= P(X_1 = 3 \mid \theta = 2) = \frac{(2)^3 \cdot \mathrm{e}^{-2}}{3!} \approx 0.18$$

The probability of an error of II kind is

$$\beta = P(X_1 \notin \mathcal{R} \mid H_1 \text{ vera}) = P(X_1 \neq 3 \mid \theta = 4)$$

$$= 1 - P(X_1 = 3 \mid \theta = 4)$$

$$= 1 - \frac{4^3 \mathrm{e}^{-4}}{3!} = 0.80$$