CE2100/CZ2100
Probability and Statistics for Computing

Tutorial 1 - Sampling Distributions

Problem 1

To determine whether a bottling machine is working satisfactorily, a production line manager randomly samples ten 12-ounce bottles every hour and measures the amount of beverage in each bottle. The mean $\bar{x}$ of the 10 fill measurements is used to decide whether to readjust the amount of beverage delivered per bottle by the filling machine.

If records show that the amount of fill per bottle is normally distributed, with a standard deviation of .2 ounce, and if the bottling machine is set to produce a mean fill per bottle of 12.1 ounces, what is the approximate probability that the sample mean $\bar{x}$ of the 10 test bottles is less than 12 ounces?

Problem 2

A college C would like to have 1050 freshmen, and cannot accommodate more than 1060. Assume that each applicant accepts with probability $p = 0.6$ and that the acceptance can be modeled with Binomial distribution. If the college accepts $\underset{n}{1700}$ freshmen, what is the probability that it will have too many acceptances?

Problem 3

The proportion of individuals with an Rh-positive blood type is 85%. You have a random sample of $n = 500$ individuals. What is the probability that the sample proportion $\hat{p}$ lies between 83% and 88%?

Problem 4

A housing survey was conducted to determine the price of a typical home in Topanga, CA. The mean price of a house was roughly $\overset{\mu}{\text{\$1.3 million}}$ with a standard deviation of \$300,000. σ
There were no houses listed below \$600,000 but a few houses above \$3 million.
What is the probability that the $\underset{\bar{x}}{\text{mean}}$ of $\underset{n}{60}$ randomly chosen houses in Topanga is more than \$1.4 million?

Problem 5

Suppose that an insurance company has $\overset{n}{10{,}000}$ policy holders. The expected yearly claim per policyholder is $\overset{\mu}{\text{\$240}}$ with a standard deviation of $\overset{\sigma}{\text{\$800}}$. What is the approximate probability that the total yearly claims $S_{10,000} >$ \$2.6 million?

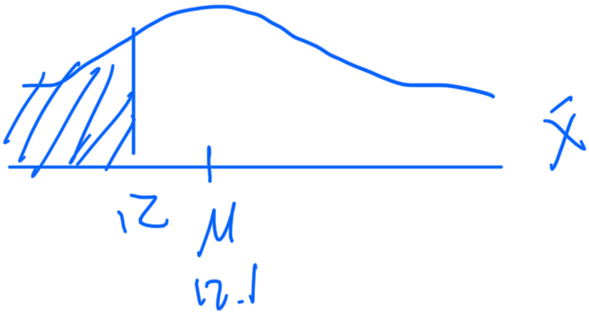
Additional Drill Questions (Do not discuss in the tutorial)

Problem 6

Suppose you roll a 6-sided die 10 times. Let X be the total value of all 10 dice, *i.e.*, $X = X_1 + X_2 + \cdots + X_{10}$. You win the game if $X \leq 25$ or $X \geq 45$. Use the central limit theorem to calculate the probability that you win.

1) $n = 10$ $\mu = 12.1$

$\bar{x}$



$P(\bar{x} < 12)$

$= P\left(Z < \dfrac{\bar{x} - \mu}{\sigma / \sqrt{n}}\right)$

$\}$

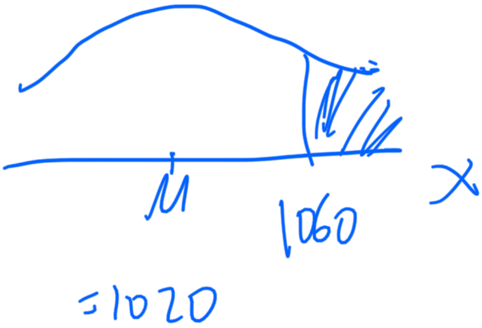
2) $\mu = n.p = 1700 . 0.6 = 1020$

$\sigma = \sqrt{npq} = 20.2$

$P(X > 1060)$

$P\left(Z > \dfrac{1060.5 - 1020}{20.2}\right)$

$\}$

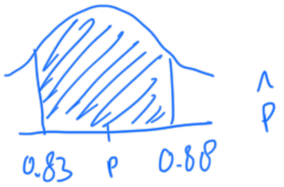


3) Sample 500 size ~ assume normal dist

$\hat{p} = \frac{X}{n}$

$E[\hat{p}] = \frac{E[X]}{n} = \frac{np}{n} = p$

$Var[\hat{p}] = \frac{Var[X]}{n^2} = \frac{npq}{n^2} = \frac{pq}{n}$



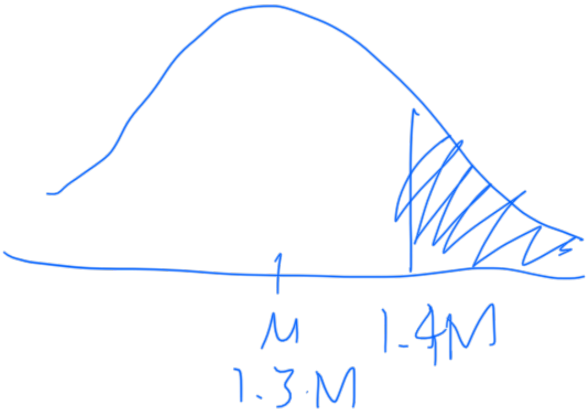
$\approx P(0.83 < \hat{p} < 0.88)$

$= P\left(\frac{0.83 - p}{\sqrt{pq/n}} < Z < \frac{0.88 - p}{\sqrt{pq/n}}\right)$

4) $\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$

$P(\bar{X} > 1.4M)$

$P\left(Z > \frac{1.4M - 1.3M}{0.3/\sqrt{60}}\right)$



5) let X = yearly claim. per policy holder

$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0,1)$

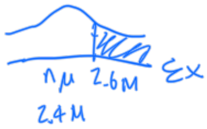
multiply by n

$\Rightarrow = \frac{n\bar{X} - n\mu}{\sigma\sqrt{n}} \sim N(0,1)$

$\Rightarrow \Sigma x \sim N(\underbrace{E[\Sigma x]}_{n\mu}, \underbrace{Var[\Sigma x]}_{n\sigma^2})$

$P(\Sigma x > 2.6M)$

$P\left(Z > \frac{2.6M - 2.4M}{\sqrt{10000} \cdot 800}\right)$



Problem 7

Suppose you have a new algorithm and want to test its running time. You have an idea of the variance of the algorithm's run time: $\sigma^2 = 4\,\text{second}^2$ but you want to estimate the mean: $\mu = t\,\text{second}$. You can run the algorithm repeatedly. How many trials do you have to run so that your estimated runtime is $t \pm 0.5$ with 95% certainty?

Problem 8

Suppose $X_1, X_2, \ldots, X_{30}$ are independent Poisson random variables with mean $\mathbb{E}(X_i) = 2$ and $Var(X_i) = 2$. Use the central limit theorem to approximate

$$P\left(\sum_{i=1}^{30} X_i > 50\right).$$

Problem 9 (Not included in quizzes)

Let X_i =weight of car i and Y_i =fuel in gallons to go 100 miles. We use the model $Y_i = \theta X_i + \epsilon_i$ where ϵ_i are independent errors with

$$\mathbb{E}[\epsilon_i] = 0, Var(\epsilon_i) = \sigma^2$$

How do we estimate θ from data? We minimize the least squares criterion

$$SS(\theta) = \sum_{i=1}^{n} (Y_i - \theta X_i)^2$$

which is minimized by

$$\hat{\theta} = \frac{\sum_{i=1}^{n} X_i Y_i}{\sum_{i=1}^{n} X_i^2}$$

What is the distribution of $\hat{\theta} - \theta$? (Note X_i is not a random variable in this question.)

CE2100/CZ2100
Probability and Statistics for Computing

Tutorial 2 - Large-Sample Estimation

Problem 1

An increase in the rate of consumer savings is frequently tied to a lack of confidence in the economy. A random sample of $n = 200$ savings accounts in a local community showed a mean increase in savings account values of 7.2% over the past 12 months, with a standard deviation of 5.6%. Estimate the mean percentage increase in savings account values over the past 12 months for depositors in the community. Find the margin of error for your estimate.

Problem 2

An opinion poll indicated that 49% of the 1034 adults surveyed think that country A should pursue a program to send humans to Mars. Estimate the true proportion of people who think that country A should pursue this program. Calculate the margin of error.

Problem 3

Each of $n = 30$ students in a chemistry class measured the amount of copper precipitated from a saturated solution of copper sulfate over a 30-minute period. The sample mean and standard deviation of the 30 measurements were equal to 0.145 and 0.0051 mole, respectively. Find a 90% confidence interval for the mean amount of copper precipitated from the solution over a 30-minute period.

Problem 4

A survey is designed to estimate the proportion of sports utility vehicles (SUVs) being driven in the state of California. A random sample of 500 registrations is selected from a Department of Motor Vehicles database, and 68 are classified as SUVs. Use a 95% confidence interval to estimate the proportion of SUVs in California.

Problem 5

It is reported that, in a study of a particular wafer inspection process, 356 dies were examined by an inspection probe and 201 of these passed the probe. Assuming a stable process, calculate a 95% (two-sided) confidence interval for the proportion of all dies that pass the probe.

Problem 6

For a study, we conduct on nutrition and access to fresh produce in Beaufort County, North Carolina. We want to know how much an adult spends on locally-produced fruit and vegetables in June. We randomly select 100 individuals from the county property records and send a survey to those residents about their eating, shopping and gardening practices. With our sample, we find that the average amount an adult spends on locally-grown fruits and vegetables in June is \$40.00. We know from previous studies that the standard deviation of money spent on local produce is \$10. Construct a 95% confidence interval for the mean (per capita) amount spent on fresh, local produce.

Additional Questions (Do not discuss in tutorial)

Problem 7

It is reported that for a sample of 50 kitchens with gas cooking appliances monitored during a one-week period, the sample mean of CO_2 level (ppm) was 654.16, and the sample standard deviation was 164.43.

(a) Calculate the 95% (two-sided) confidence interval for the true average CO_2 level in the population of all homes from which the sample was selected.

(b) Suppose the investigators had made a rough guess of 175 for the value of standard deviation s before collecting data. What sample size would be necessary to obtain an interval width of 50 ppm for a confidence level of 95%?

Problem 8

Let $X_1, \ldots, X_n$ be a random sample from population with mean μ and variance σ^2. Show that the sample variance $S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2$ is an unbiased estimator for σ^2.

CE2100/CZ2100
Probability and Statistics for Computing

Tutorial 3 - Large-Sample Tests of Hypotheses

In the following questions, it is assumed that the sample sizes are large enough for applying central limit theorem.

Problem 1

A four-sided (tetrahedral) die is tossed 1000 times, and 290 fours are observed. Is there evidence to conclude that the die is biased, that is, say, that more fours than expected are observed? (Use $\alpha = 0.05$)

Problem 2

Let p equal the proportion of drivers who use a seat belt in a state that does not have a mandatory seat belt law. It was claimed that $p = 0.14$. An advertising campaign was conducted to increase this proportion. Two months after the campaign, $y = 104$ out of a random sample of $n = 590$ drivers were wearing seat belts. Was the campaign successful? (Use $\alpha = 0.01$)

Problem 3

Boys of a certain age are known to have a mean weight of $\mu = 85$ pounds. A complaint is made that the boys living in a municipal children's home are underfed. As one bit of evidence, $n = 25$ boys (of the same age) are weighed and found to have a mean weight of $\bar{x} = 80.94$ pounds. It is known that the population standard deviation σ is 11.6 pounds. Based on the available data, what should be concluded concerning the complaint using the p-value approach? It is assumed that weight follows a normal distribution. (Use $\alpha = 0.05$)

Problem 4

To compare customer satisfaction levels of two competing cable television companies, 174 customers of Company 1 and 355 customers of Company 2 were randomly selected and were asked to rate their cable companies on a five-point scale, with 1 being least satisfied and 5 most satisfied. The survey results are summarized in the following table

Company 1	Company 2
$n_1 = 174$	$n_2 = 355$
$\bar{x}_1 = 3.51$	$\bar{x}_2 = 3.24$
$s_1 = 0.51$	$s_2 = 0.52$

$H_0 = \mu_1 - \mu_2 7 0$

$H_a = \mu_1 - \mu_2 > 0$

Test at the 1% level of significance whether the data provide sufficient evidence to conclude that Company 1 has a higher mean satisfaction rating than Company 2.

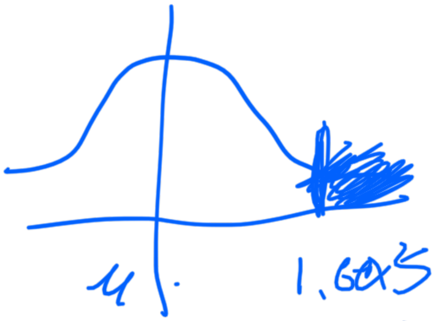
Additional Questions (Do not discuss in the tutorial)

Problem 5

The melting point of each of 16 samples of a certain brand of hydrogenated vegetable oil

1) $H_0 : \frac{1}{4} = 0.25$ $H_0 = P = 0.25$ $Z = 1.645$

$H_a : p > 0.25$



$$P = \frac{290}{1000} = 0.29$$

$$Z = \frac{0.29 - 0.25}{\sqrt{pq}/\sqrt{n}} = \frac{0.04}{\sqrt{(0.29)(0.75)/1000}}$$

$= 2.92 > 1.645$

reject H_0

1.645	$\alpha/2 = 0.05$
1.960	$\alpha/2 = 0.025$
2.326	$\alpha/2 = 0.01$
0.675	$\alpha/2$ 0.25

2) H_0 $p = 0.14$ H_a $P > 0.14$ $Z = 0.01$

$\to Z = 2.326$

$$P = \frac{104}{590} = 0.18$$

$$Z = \frac{0.18 - 0.14}{\sqrt{0.14 \; 0.86 / 590}} = \frac{0.04}{0.0143} = 2.8$$

$2.8 > 2.326$

H_0 reject

3) $\mu = 85$ $\quad H_0 : \mu = 85 \quad H_a : \mu < 85$

$80.94 \quad 25$

$\sigma = 11.6$

$$\frac{80.94 - 85}{11.6/\sqrt{25}} = 1.75$$

$$P(Z > 1.75) = (1 - 0.9599) = \underline{0.0401} < 0.05$$

$1.75 > 1.645$ $\quad$ Ho is false

was determined, resulting in $\bar{x} = 94.32$. Assume that the distribution of the melting point is normal with $\sigma = 1.20$. Test $H_0 : \mu = 95$ versus $H_a : \mu \neq 95$ using a two-tailed level 0.01 test.

Problem 6

The desired percentage of SiO_2 in a certain type of aluminous cement is 5.5. To test whether the true average percentage is 5.5 for a particular production facility, 16 independently obtained samples are analyzed. Suppose that the percentage of SiO_2 in a sample is normally distributed with $\sigma = 0.3$ and that $\bar{x} = 5.25$. Assuming that 16 samples are enough to apply CTL, does the above information indicate conclusively that the true average percentage differs from 5.5 under common significance levels?

Problem 7

A statistical statement appeared in *The Guardian* on Friday January 4, 2002:

> When spun on edge 250 times, a Belgian one-euro coin came up heads 140 times and tails 110. 'It looks very suspicious to me', said Barry Blight, a statistics lecturer at the London School of Economics. 'If the coin were unbiased the chance of getting a result as extreme as that would be less than 6%'.

(a) Let θ be the probability of coming up heads. Consider the null hypothesis that the coin is fair, $H_0 : \theta = 0.5$. Carefully explain how the 6% figure arises. What term describes this value in hypothesis testing? Does it correspond to a one-sided or two-sided test?

(b) Would you reject H_0 at a significance level of $\alpha = 0.1$? What about $\alpha = 0.05$? (Using two-sided test)

(c) How many heads would you need to have observed out of 250 spins to reject at a significance of $\alpha = 0.01$?

CE2100/CZ2100
Probability and Statistics for Computing

Tutorial 4 - Inference from Small Samples

Problem 1

A small component in an electronic device has two small holes where another tiny part is fitted. In the manufacturing process, the average distance between the two holes must be tightly controlled at 0.02 mm, else many units would be defective and wasted. Many times throughout the day quality control engineers take a small sample of the components from the production line, measure the distance between the two holes, and make adjustments if needed. Suppose at one time four units are taken and the distances are measured as

$$0.021 \quad 0.019 \quad 0.023 \quad 0.020$$

Determine, at the 1% level of significance, if there is sufficient evidence in the sample to conclude that an adjustment is needed. Assume the distances of interest are normally distributed.

Problem 2

A state agency requires a minimum of 5 parts per million (ppm) of dissolved oxygen in order for the oxygen content to be sufficient to support aquatic life. Six water specimens taken from a river at a specific location during the low-water season (July) gave readings of 4.9, 5.1, 4.9, 5.0, 5.0 and 4.7 ppm of dissolved oxygen. Do the data provide sufficient evidence to indicate that the dissolved oxygen content is less than 5 ppm? Test using $\alpha = 0.05$.

Problem 3

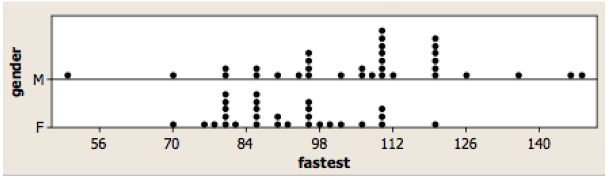
A psychologist was interested in exploring whether or not male and female college students have different driving behaviours. There were a number of ways that she could quantify driving behaviours. She opted to focus on the fastest speed ever driven by an individual. Therefore, the particular statistical question she framed was as follows:

Is the mean fastest speed driven by male college students different from the mean fastest speed driven by female college students?

She conducted a survey of a random $n = 34$ male college students and a random $m = 29$ female college students. Here is a descriptive summary of the results of her survey:

Males (X)	Females (Y)
$n = 34$	$m = 29$
$\bar{x} = 105.5$	$\bar{y} = 90.9$
$s_x = 20.1$	$s_y = 12.2$

and here is a graphical summary of the data in the form of a dotplot:



Is there sufficient evidence at the $\alpha = 0.05$ level to conclude that the mean fastest speed driven by male college students differs from the mean fastest speed driven by female college students?

Problem 4

In a packing plant, a machine packs cartons with jars. It is supposed that a new machine will pack faster on average than the machine currently used. To test that hypothesis, the times it takes each machine to pack ten cartons are recorded. The results, in seconds, are shown in the tables.

New machine	Old machine
42.1	42.7
41	43.6
41.3	43.8
41.8	43.3
42.4	42.5
42.8	43.5
43.2	43.1
42.3	41.7
41.8	44
42.7	44.1

Do the data provide sufficient evidence to conclude that, on average, the new machine packs faster?

Problem 5

In a study on the effect of an oral rinse on plaque buildup on teeth, 14 people whose teeth were thoroughly cleaned and polished were randomly assigned to two groups of seven subjects each. Both groups were assigned to use oral rinses (no brushing) for a 2-week period. Group 1 used a rinse that contained an antiplaque agent. Group 2, the control group, received a similar rinse except that the rinse contained no antiplaque agent. A measure of plaque buildup was recorded at 14 days with means and standard deviations for the two groups shown in the table

	Control Group	Antiplaque Group
Sample Size	7	7
Mean	1.26	0.78
Standard Deviation	0.32	0.32

Do the data provide sufficient evidence to indicate that the oral antiplaque is effective? Test using $\alpha = 0.05$.

Additional Questions (Not discuss in tutorial)

Problem 6

We perform a t-test for the null hypothesis $H_0 : \mu = 10$ at significance level $\alpha = 0.05$ by means of a dataset consisting of $n = 16$ elements with sample mean 11 and sample variance 4. Use p-value to solve the following questions.

(a) Should we reject the null hypothesis in favour of $H_a : \mu \neq 10$?

(b) What if we test against $H_a : \mu > 10$?

Problem 7

Suppose μ is the average height of a college male. You measure the heights (in inches) of twenty college men, getting data $x_1, \ldots, x_{20}$, with sample mean $\bar{x} = 69.55$ in. and sample variance $s^2 = 14.26$ in^2. Suppose that the x_i are drawn from a normal distribution with unknown mean μ and unknown variance σ^2.

(a) Using $\bar{x}$ and s^2, construct a 90% t-confidence interval for μ.

(b) Now suppose you are told that the height of a college male is normally distributed with a standard deviation 3.77 in. Construct a 90% z-confidence interval for μ.

(c) In (b), how many people in total would you need to measure to bring the width of the 90% z-confidence interval down to 1 inch?

(d) Consider again the case of unknown variance in (a). Based on this sample variance of 14.26 in^2, how many people in total should you expect to need to measure to bring the width of the 90% t-confidence interval down to 1 inch? Is it guaranteed that this number will be sufficient? Explain your reasoning.

Problem 8

Consider a machine that is known to fill soda cans with amounts that follow a normal distribution with (unknown) mean μ and standard deviation $\sigma = 3$ mL. We measure the volume of soda in a sample of bottles and obtain the following data (in mL):

$$352, 351, 361, 353, 352, 358, 360, 358, 359$$

(a) Construct a precise 95% confidence interval for the mean μ

(b) Now construct a 98% confidence interval for the mean μ

(c) Suppose now that σ is not known. Redo parts (a) and (b), and compare your answers to those above.

Problem 9

The following data comes from a real study in which 1408 women were admitted to a maternity hospital for (i) medical reasons or through (ii) unbooked emergency admission. The duration of pregnancy is measured in complete weeks from the beginning of the last menstrual period. We can summarize the data as follows:
Medical: 775 observations with $\bar{x}_M = 39.08$ and $s_M^2 = 7.77$.
Emergency: 633 observations with $\bar{x}_E = 39.60$ and $s_E^2 = 4.95$
Set up and run a two-sample t-test using p-value to investigate whether the mean duration differs for the two groups. What assumptions did you make?