Chapter 1

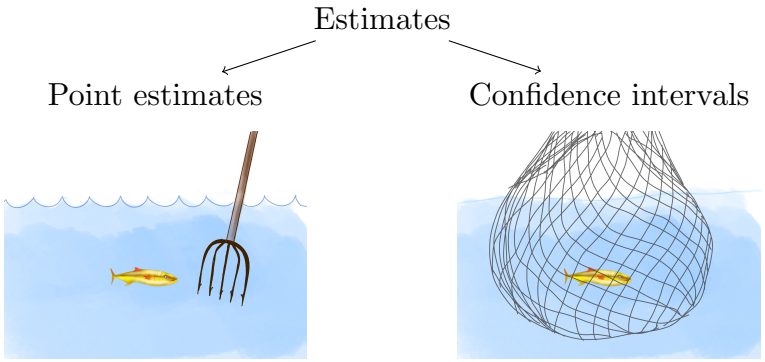
Confidence Intervals

> Confidence comes not from always being right, but from not fearing being wrong.
>
> Peter T. McIntyre

Point estimation versus interval estimation

As it was discussed earlier, we are usually interested in finding population parameters. Examples of parameters are: μ – the mean salary in Russia in 2017, p – the population proportion of Saint-Petersburg citizens in favor of the anti-gay law, or σ – true riskiness of some investment project. The problem is that we cannot *calculate* those parameters as we do not have the data about the whole population. However, we can *estimate* their values based on the sample data.

One way to do this is to calculate **point estimates**. As we have shown in Chapter 9, sample mean ($\bar{X}$), sample proportion ($\hat{p}$) and sample standard deviation s can be used as approximations for the corresponding parameters. This approach has its limitations, as point estimates almost never coincide with the exact true values of parameters. To increase investigator's confidence in estimates the other approach is used – **interval estimates**. Although a point estimate "almost never" catches the value of true parameter, it can be used to construct an interval which will include the parameter "almost for sure".



You may compare these two methods with the two approaches to fishing. Trying to catch a fish with a *spear* is similar to calculating *point* estimate in hope it will be equal to the parameter, only a few attempts will be successful. Contrary, using a *net* is similar to constructing an *interval*, which would contain the parameter in significant number of trials.

In each particular case you should decide which method is currently more appropriate. Each of them has its advantages and drawbacks. Point estimation gives a single number as an estimate, which is very convenient for further analysis. However, it usually fails to "catch" the true value. Interval provides a researcher with higher confidence in her estimate of the parameter, but it might be inconvenient to use in further analysis.

The discussed interval estimates are called confidence intervals since they contain the true parameter with a specified level of confidence.

How to construct a confidence interval?

Let us suppose we have a big sample of values, obtained by SRS, of some variable X. According to the Central Limit Theorem (CLT) sample mean $\bar{X}$ has normal distribution (since n is large). As we've learned from the previous chapter, it has the following parameters: $\bar{X} \sim N(\mu, \frac{\sigma}{\sqrt{n}})$.

Step 1. As for any Normal variable we can reduce $\bar{X}$ to the Standard Normal variable z:

$$\frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} = z \sim N(0, 1) \tag{1.1}$$

Step 2. If we express $\bar{X}$ from the equation (1) we get: $\bar{X} = \mu + z \cdot \frac{\sigma}{\sqrt{n}}$. z is a random variable taking positive as well as negative values. Thus, the first summand shows that $\bar{X}$ on average equals μ, and the second summand represents the deviation of $\bar{X}$ from μ which depends on a random component z. z follows standard normal distribution and answers the question: "By how many standard deviations ($\frac{\sigma}{\sqrt{n}}$) did $\bar{X}$ fall away from μ?". So, the equation specifies the range of the deviations of $\bar{X}$ around μ which might happen with given probability reflected in the corresponding value of z. Denoting z-critical as the absolute (positive) value of z we can rewrite the expression as follows:

$$\bar{X} = \mu \pm z_{\text{critical}} \cdot \frac{\sigma}{\sqrt{n}} \tag{1.2}$$

Step 3. However, we are usually interested in plausible values of μ, not $\bar{X}$. Expressing μ from the equation (2) we get:

$$\mu = \bar{X} \pm z_{\text{critical}} \cdot \frac{\sigma}{\sqrt{n}}$$

Thus, we've arrived to the frame for the first formula of confidence interval. It provides a convenient way to estimate boundaries within which the true value of mean μ lies with certain level of confidence.

The component z-critical $\times \frac{\sigma}{\sqrt{n}}$ is called the **margin of error** (sometimes abbreviated as MoE).

Analogically, you can derive a formula for any other parameter. Any confidence interval will match to the following structure:

$$\text{parameter} = \text{estimator} \pm \text{MoE}$$

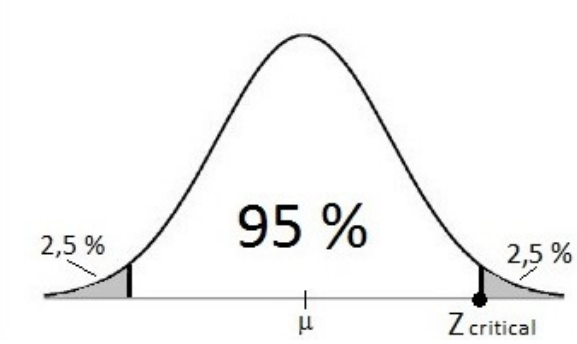
$$\text{MoE} = \text{critical statistic} \times \text{standard error of the estimator}$$

What is Margin of Error and how to find z-critical?

Masha is interested in how tall the average ICEF student is. Let X be the height of an ICEF student. Assume the true mean is $\mu = 175$cm and standard deviation is $\sigma = 5$cm. A simple random sample of 100 students is taken and their heights are measured. The sample mean height $\bar{X}$ can take different values depending on heights of random set of students, who get into the sample. By CLT $\bar{X}$ has approximately normal distribution: $\bar{X} \sim N(175, \frac{5}{\sqrt{100}})$. Based on the equation (2) the range of its possible values is: $\bar{X} = 175 \pm z\text{-critical} \times \frac{5}{\sqrt{100}}$

Suppose we want the interval to include 95% of values $\bar{X}$ can take. What value of z-critical should be used?

The picture below displays probability distribution of $\bar{X}$. In order to find the smallest interval which includes the most probable 95% of values, we should take the central 95% of distribution. This leaves aside 5% of unlikely values at the two tails of the distribution:



Due to the symmetry, area of each tail equals $\frac{5\%}{2} = 2.5\%$. Thus, z-critical $= z_{0.025}$, such that: $P(z > z_{0.025}) = 2.5\% = 0.025$. From the table we get: $z_{0.025} \approx 1.96$.

Then, $\bar{X} = 175 \pm 1.96 \cdot \frac{5}{\sqrt{100}}$. So, we get that with 95% probability the sample mean will take value between 174 and 176 centimeters.

Therefore, in general case z-critical $= z_{\frac{\alpha}{2}}$ and $\bar{X}$ takes values within the range $\mu \pm z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}$ with $1 - \alpha$ probability.

Reversing the equation to express μ we get the general formula for confidence interval for mean:

$$\mu = \bar{X} \pm z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}} \tag{1.3}$$

The value $(1-\alpha)$ is called **confidence level**. It expresses the probability that confidence interval will include the true value of parameter. In the given example confidence level is 95 percent.

Contrary, **significance level** α is the probability that confidence interval will not include the true parameter.

Remember, that there is always a chance that confidence interval *will not* contain the true value of parameter!

Confidence intervals for population parameters (μ and p)

Confidence intervals for the population mean (μ)

The case of known σ

As we have shown above, the formula for confidence interval for μ given σ is known is:

$$\mu = \bar{X} \pm z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}$$

Is there life without the Central Limit Theorem? We've started our derivation of the formula by stating that for large enough sample the CLT insures that: $\bar{X} \sim N(\mu, \frac{\sigma}{\sqrt{n}})$. However, what happens if the sample is small, say, less than 30 observations? Does that mean that construction of confidence interval is impossible? The answer depends on the distribution of X. Note that if X is itself normally distributed sample mean $\bar{X}$, being a sum of independent normal variables is also normal. Thus, we can apply the same formula even for small samples if $X \sim N$. If this is not true – the standard formula for confidence interval cannot be used.

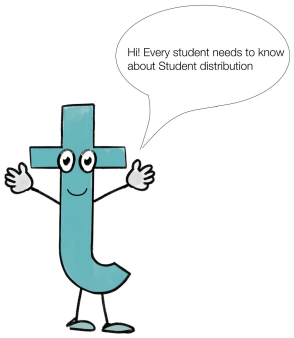
So, if you are asked to find the confidence interval based on the small sample, and nothing is said about the distribution of X, *be sure to state the assumption* that the random variable of interest is normally distributed before you estimate the interval. If you are given sample data then you must check if this assumptions is plausible: look at the shape of the distribution (it should be symmetric and bell-shaped), check if there are outliers (in the case of normal distribution it is highly unlikely to observe an outlier in a small sample). If there are no obvious departures from normality of X, you can state this assumption and proceed in calculating the interval.

Confidence interval for μ with unknown σ

So far we've reached the formula: $\mu = \bar{X} \pm Z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}$ to estimate the unknown population mean. To calculate such an interval the value of σ needs to be known. However, when population mean is unknown (we are trying to estimate it by the interval!) how is it possible to know the true value of population standard deviation? Indeed, it is usually not the case. All we have is a sample! Of course, we can replace σ with its sample estimate s.

As we've shown earlier in this section, $\frac{\bar{X}-\mu}{\frac{\sigma}{\sqrt{n}}} \sim N(0,1)$. How the probability distribution will change if σ is replaced with s?

Obviously, when a constant is replaced with a random variable the variance of the whole expression will rise. Pdf of a random variable $\frac{\bar{X}-\mu}{\frac{s}{\sqrt{n}}}$ will have bigger "tails" indicating higher probability of deviations from mean. What is more, the smaller is the sample – the higher is the probability of large deviations of s from σ, reflected in even fatter tails. It happens because small samples give less precise values of standard deviation s, which becomes more dispersed.



However, all the other properties of pdf function stay the same: centered at 0 and having bell-shaped form. Here the new distribution is to be introduced – the **Student or t-distribution**. Its pdf is defined by the so-called degrees of freedom (df) calculated as number of observation minus 1, df $= n - 1$. The corresponding number is conventionally indicated in brackets after the t letter: $t(n-1)$. The closer is s to σ (which in general happens with higher n), the closer is t-distribution to z.

As is shown on the picture above t-distributions with lower degrees of freedom are more dispersed around 0 (v on the picture denotes value of df). On the other hand, with larger df t-distribution becomes closer to $N(0,1)$. Indeed, when n approaching infinity, pdf of t approaches pdf of z.

Thus, we get the following formula:

$$\frac{\bar{X}-\mu}{\frac{s}{\sqrt{n}}} \sim t(n-1)$$

Applying the same algorithm as in 1.1 we arrive to the new formula of confidence interval:

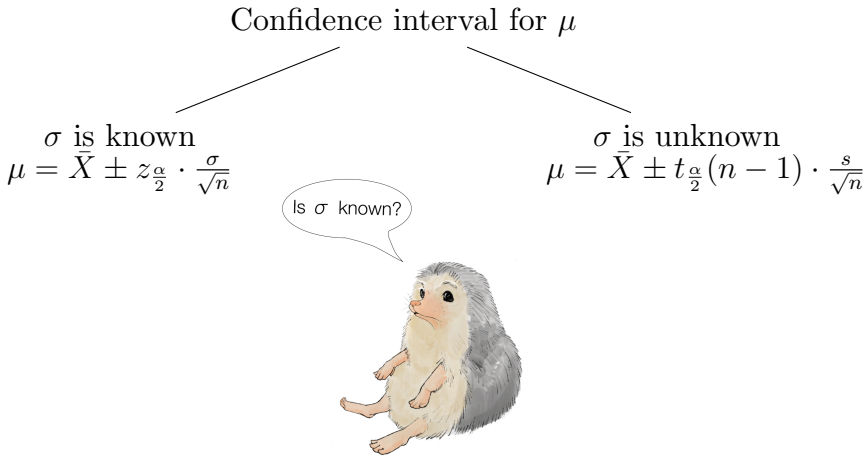
$$\mu = \bar{X} \pm t_{\frac{\alpha}{2}}(n-1) \cdot \frac{s}{\sqrt{n}}$$

Note that $t_{\frac{\alpha}{2}}(n-1)$ is a number (not multiplication of t by $(n-1)$)! It is a critical value of t-distribution with $n-1$ degrees of freedom such that $P(t > t_{\frac{\alpha}{2}}(n-1)) = \frac{\alpha}{2}$.

Note that when n is large enough t-distribution can be approximated by z. Thus, for large samples it is acceptable to apply the formula: $\mu = \bar{X} \pm z_{\frac{\alpha}{2}} \cdot \frac{s}{\sqrt{n}}$. However, it would be an approximation, while the general rule is to use z-distribution when σ is known, and t-distribution otherwise.

Again, the formulas given here are applicable when sample is large enough so that we can apply CLT. For small samples to ensure that $\bar{X} \sim N$ we need to make sure

that X is normal. If it's not given in a problem – you should state the corresponding assumption.



Confidence interval for population proportion p

Analogically, by CLT for large n sample proportion p has normal distribution. As we've shown in the previous chapter $\hat{p} \sim N(p, \sqrt{\frac{p(1-p)}{n}})$

Thus, $z = \frac{\hat{p}-p}{\sqrt{\frac{p(1-p)}{n}}}$ and $p = \hat{p} \pm Z_{\frac{\alpha}{2}} \cdot \sqrt{\frac{p(1-p)}{n}}$

Now $z_{\frac{\alpha}{2}} \cdot \sqrt{\frac{p(1-p)}{n}}$ is the margin of error. Since we do not know the true value of proportion p we cannot directly calculate the margin of error. However, as the sample is initially assumed to be large enough we can still construct an approximate confidence interval by replacing p in margin of error with its sample estimate $\hat{p}$.

Thus, we get the conventional formula for the confidence interval for p:

$$p = \hat{p} \pm z_{\frac{\alpha}{2}} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

Note that *this is the only formula* used for (2-sided) interval estimation of p. It is only used when sample is large enough and uses only normal distribution (no t-distribution for proportions!).

Confidence intervals for the difference of parameters

Confidence interval for difference in population means ($\mu_1 - \mu_2$)

1. Known population standard deviations (σ_1, σ_2)

As we've shown in Chapter 8, under certain conditions difference of two sample means has the following normal distribution: $\bar{X}_1 - \bar{X}_2 \sim N\left(\mu_1 - \mu_2, \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}\right)$.

Recall these conditions: samples of both X_1, X_2 are SRS, $n_1 \geqslant 30$, $n_2 \geqslant 30$ and X_1 and X_2 are independent from each other. Alternatively, if n is small, you should also check that $\bar{X}_1$ and $\bar{X}_2$ are both normal. Thus,

$$\frac{\bar{X}_1 - \bar{X}_2 - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} = z \sim N(0,1)$$

Hence, the following confidence interval is constructed:

$$\mu_1 - \mu_2 = \bar{X}_1 - \bar{X}_2 \pm z_{\frac{\alpha}{2}} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

2. Unknown population standard deviations, when $\sigma_1 \neq \sigma_2$

Of course the previous formula can only be used given you know the true values of standard deviations σ_1 and σ_2 of X_1 and X_2. In most real life problems this is not the case. All we usually have are the samples, and thus, the sample statistics calculated on them. When we replace σ_1 and σ_2 with sample standard deviations s_1 and s_2 the distribution is no more normal. To account for the increased variability in possible values of $\bar{X}_1 - \bar{X}_2$ Student distribution is applied.

$\frac{\bar{X}_1 - \bar{X}_2 - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \sim t\,(k-1)$, $k = \min\{n_1, n_2\}$

Thus, we come to the following formula of confidence interval:

$$\mu_1 - \mu_2 = \bar{X}_1 - \bar{X}_2 \pm t_{\frac{\alpha}{2}}(k-1) \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}},\ k = \min\{n_1, n_2\}$$

Why is the minimum of n_1 and n_2 is used in calculating the degrees of freedom? The answer is that variability of the whole construction depends on the "worst", the least stable component in it. Let's suppose $n_1 = 1000$, $n_2 = 25$. That means calculations of s_1 are highly precise, it takes values close to the true standard deviation σ_1 and has small variance. Contrary, s_2, being calculated on the small sample is highly volatile. Hence, however precise s_1 is, variability in possible values of s_2 will make the overall expression less stable, leading to small degrees of freedom – depending on n_2. Hence, degrees of freedom are always calculated based on the size of the smallest sample.

Note that your graphic calculator uses another and somewhat more sophisticated calculation of degrees of freedom. It results in a bit higher and usually fractional (not integer) value of degrees of freedom. If you use the results from the calculator in solving a problem you should indicate df used by it.

If the samples are not large enough (in the AP course sample is supposed to be small when $n < 30$) you can still apply the same formula given that both X_1 and X_2 are normal random variables.

3. Unknown population standard deviations assuming $\sigma_1 = \sigma_2$

In many situations it is reasonable to assume that $\sigma_1 = \sigma_2 = \sigma$. What happens then?

$$\mu_1 - \mu_2 = \bar{X}_1 - \bar{X}_2 \pm z_{\frac{\alpha}{2}} \cdot \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} = \bar{X}_1 - \bar{X}_2 \pm z_{\frac{\alpha}{2}} \cdot \sigma \cdot \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

The above formula can be applied when σ is known.

When it's unknown σ can be evaluated based on the joint sample of X_1 and X_2. Of course, since population standard deviations for both X_1 and X_2 are the same (σ) you can also estimate it based on either X_1 sample (use s_1) or X_2 sample (use s_2). However, the preciseness of the both estimators will be limited to corresponding sample size (n_1 or n_2). To make the estimator more precise it is useful to merge the two samples.

The generally used estimator for σ is the so called **pooled standard deviation** (for the explanation of formula address the Appendix):

$$\hat{\sigma} = s_p = \sqrt{\frac{s_1^2 \cdot (n_1 - 1) + s_2^2 \cdot (n_2 - 1)}{n_1 + n_2 - 2}}$$

When we replace σ with s_p normal distribution changes to Student distribution. Since the preciseness of s_p is limited to the size of joint sample $n_1 + n_2$ and its calculation involves two estimates of unknown population parameters Student distribution has $(n_1 + n_2 - 2)$ degrees of freedom.

Thus, we arrive to the formula for confidence interval with the same and unknown standard deviation σ:

$$\mu_1 - \mu_2 = \bar{X}_1 - \bar{X}_2 \pm t_{\frac{\alpha}{2}}(n_1 + n_2 - 2) \cdot s_p \cdot \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

4. Matched samples case

Sometimes the difference is to be calculated based on so-called matched or paired samples. It means that although two sets of observations are involved (say, X_1 and X_2), both are in fact taken from the same population of X.

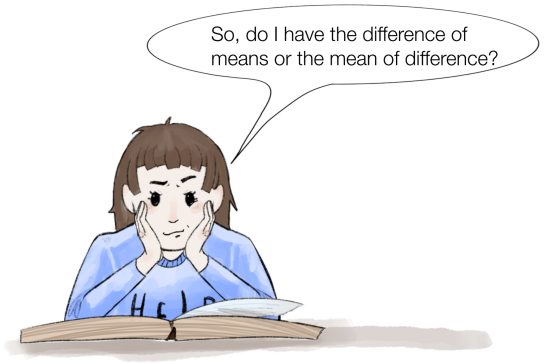
For example you may want to compare the mean difference in blood pressure of 1st year ICEF students before and after they take their winter exam in Stats (population is presented by all ICEF freshmen). For that purpose one sample of 1st year students can be taken and their average measures of blood pressure before and after the exam ($\bar{X}_1$ and $\bar{X}_2$) are used to construct the interval for mean difference (we denote the true mean difference by μ_{diff}). In that case not only the population is the same (1st year students), but the samples contain the same set of students. Why we compare a sample with itself? Because we want to see is there effect of some treatment which is not applied to sample objects at the first stage (we observe X_1) and is applied at the second stage (X_2). In the given example treatment is the exam and we want to see does it constitute a stress for students. In order to check that we look whether the confidence interval for μ_{diff}, estimated by $\bar{d} = (\overline{X_1 - X_2})$, contains zero or not. If it does, then, $X_{1i} - X_{2i}$ (difference in blood pressure of student i) on average is close to zero and the interval of most plausible values of mean difference includes the case of no effect (zero). Therefore, we conclude that the treatment has no statistically measurable effect. Contrary, if the mean difference interval only contains positive (or

only negative) values we may assert that the exam is a stressful event for students since it significantly changes their physiological condition.

However, sometimes it's impossible to get exactly the same samples, and the two samples from the same population are taken. For example we might have a sample of twins, who has almost the same DNA. But in each pair of twins the first one was brought up by the mother only (sample 1), while the second was living with the father (sample 2). In the given example we might be interested in effects of gender of a single parent on psychological characteristics of a child. Say, we may want to compare mean difference in feminity scores shown up by the twins. In this example population is the same (all twins, one of whom was brought up by mother, and the other – by father). Samples are different, but can be easily paired, so that we may compare the scores of twins in each pair. The strategy for constructing the confidence interval is the same. First, we create one sample from the two – we calculate differences in scores for each pair i: $d_i = X_{1i} - X_{2i}$. Then, we apply the same strategy as for constructing confidence interval for single mean μ. We calculate sample mean $\bar{d}$ and sample standard deviation s_d and apply t-distribution to estimate μ_{diff}:

$$\mu_{\text{diff}} = \bar{d} \pm t_{\frac{\alpha}{2}}(n-1) \cdot \frac{s_d}{\sqrt{n}}$$

As usually, if sample of pairs is small ($n < 30$), we should check (or at least assume) that d is normally distributed.



Confidence interval for $\mu_1 - \mu_2$

↓

Are the samples matched/paired?

→ Yes, they are taken from same population X

$$\mu_{\text{diff}} = \bar{d} \pm t_{\frac{\alpha}{2}}(n-1) \cdot \frac{s_d}{\sqrt{n}}$$

No, they refer to different populations of X_1 and X_2

→ σ_1 and σ_2 are known

$$\mu_1 - \mu_2 = \bar{X}_1 - \bar{X}_2 \pm z_{\frac{\alpha}{2}} \cdot \sqrt{\frac{\sigma_1}{n_1} + \frac{\sigma_2}{n_2}}$$

σ_1 and σ_2 are unknown

→ $\sigma_1 \neq \sigma_2$

$$\mu_1 - \mu_2 = \bar{X}_1 - \bar{X}_2 \pm t_{\frac{\alpha}{2}}(\text{df}) \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

$$\text{df} = k - 1 = \min\{n_1, n_2\} - 1$$

for calculating manually
using the calculator, take value provided there

$\sigma_1 = \sigma_2$

$$\mu_1 - \mu_2 = \bar{X}_1 - \bar{X}_2 \pm t_{\frac{\alpha}{2}}(\text{df}) \cdot s_p \cdot \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

$$\text{df} = n_1 + n_2 - 2$$

Note that none of these formulas can be applied unless all the assumptions hold. Using formulas without checking assumptions is like puching accelerator when you forgot to start the engine in the car. It simply does not work! All the assumptions applied in for the confidence interval estimation are provided at the end of the chapter. Be sure to learn them!

Dear Mu, don't be so stressed! These are all modifications of one and the same formula:

$$\text{parameter} = \text{estimate} \pm \text{MoE}$$

$$\text{MoE} = \text{critical statistic} \times \text{standard error of the estimator}$$

If you have understanding of the confidence interval mechanics you don't need to remember any of the formulas by heart. On the exam you'll be given a table with formulas for standard deviations for all the cases concerned.

Confidence interval for the difference in populations proportions $p_1 - p_2$

It was previously shown that difference in population proportions is normally distributed given that both samples are large enough:

$$\hat{p}_1 - \hat{p}_2 \sim N(p_1 - p_2, \sqrt{\frac{p_1 \cdot (1 - p_1)}{n_1} + \frac{p_2 \cdot (1 - p_2)}{n_2}})$$

Thus, we have:

$$\frac{\hat{p}_1 - \hat{p}_2 - (p_1 - p_2)}{\sqrt{\frac{p_1 \cdot (1-p_1)}{n_1} + \frac{p_2 \cdot (1-p_2)}{n_2}}} = z \sim N(0, 1)$$

Since we do not know the true values of proportions p_1 and p_2 we cannot directly calculate the margin of error, we only approximate it by replacing p_1 and p_2 with $\hat{p}_1$ and $\hat{p}_2$. Hence, the following confidence interval is constructed:

$$p_1 - p_2 = \hat{p}_1 - \hat{p}_2 \pm z_{\frac{\alpha}{2}} \cdot \sqrt{\frac{\hat{p}_1 \cdot (1 - \hat{p}_1)}{n_1} + \frac{\hat{p}_2 \cdot (1 - \hat{p}_2)}{n_2}}$$

Note that this is the only formula used for (2-sided) interval estimation of $p_1 - p_2$. It is only used when samples are large enough and uses only normal distribution (no t-distribution for proportions!).

One-sided confidence interval

Before she started her studies at ICEF Masha used to live in some Russian city far away from the Moscow region. Chuvyakin is a governor of that city. Masha has read in the news that he stated that the average monthly salary in the city is 55 000 roubles. She is suspicious about this statement, since, based on her experience, the real figure should be much lower.

Let's denote salary of a citizen by X, the true mean salary by μ.

Masha has found that (based on previously gathered data) population standard deviation is known to be $\sigma = 15\ 500$ roubles. Masha took a simple random sample of 40 citizens, obtained data on their salaries X, and calculated average salary to be $\bar{X} = 35\ 000$, which is much lower than 55 000 – the estimate provided by Chuvyakin.

Is it because of the sampling error or the true mean salary is indeed lower than what Chuvyakin has stated?

Masha decided to answer this question.

Ok, we have a simple random sample of size $n = 40 > 30$. So, This is enough to state that $\bar{X} \sim N(\mu, \frac{\sigma}{\sqrt{n}})$. Based on the sample data she's obtained, what is the

largest possible result that a random sample, assumingly taken by Chuvyakin team, could reasonably result in? Say, among all possible values of sample mean he could get, what is the value, such that 99% of all other samples does not exceed? Let's denote this value by $\bar{X}_{0.99}$ meaning that it is the maximum plausible value of the mean at $p = 99\%$. In other words, what is the 99-percentile for the distribution of $\bar{X}$?

Using the same logic as on pp.2-3 you can get that: $\mu < \bar{X} + z_{0.1}\frac{\sigma}{\sqrt{n}}$ at 99% confidence level.

So, we are 99% sure that $\mu < 35\ 000 + 1,282\frac{15\ 500}{\sqrt{40}} \approx 38\ 142$ roubles. Based on that result, there is enough statistical evidence to state that Chuvyakin severely overestimates the true mean salary.

The general formulas for one-sided confidence intervals are:

$$\mu < \bar{X} + z_\alpha \frac{\sigma}{\sqrt{n}}$$

$$\mu > \bar{X} - z_\alpha \frac{\sigma}{\sqrt{n}}$$

Conditions/Assumption for confidence intervals

Assumption for confidence intervals on population mean/difference in means

1. **The sample is SRS**. Don't forget to check/assume that sample is less than 5% of the whole population.
2. **Normality** of your random variable:
   - if n is large ($n > 30$)
   - n is small – check/assume that X (or X_1 and X_2) is normal (symmetry, no outliers)
3. (for $\mu_1 - \mu_2$) Independent samples

Assumption for confidence intervals on population proportion/difference in proportions

1. The sample is SRS.
2. Normality of your random variable (if $np > 5$ and $n(1-p) > 5$ it is true by CLT)
3. (for $p_1 - p_2$) Independent samples

You must MUST BE ABLE TO REPRODUCE even being half-awake

- parameter = estimate ± MoE,
  MoE = critical statistic × standard error of the estimator (structure of any confidence interval)
- Which formula to apply in each case. There is no need to learn them by heart if you remember the structure of any confidence interval and know when to apply z- and t-statistics. Standard deviations of $\bar{X}$, $\hat{\rho}$, $\bar{X}_1 - \bar{X}_2$, $\hat{\rho}_1 - \hat{\rho}_2$ necessary in the formulas of CI are present in the AP formulas sheet.
- Assumptions and conditions to check/state for each of the formulas to apply.
- How to interpret the interval

Calculator Box

- $\mu = \bar{X} \pm z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}$: `INTR` → `z` → `1-s` (stands for 1 sample) → enter values of statistics → `Exe`
- $\mu = \bar{X} \pm t_{\frac{\alpha}{2}}(n-1) \cdot \frac{s}{\sqrt{n}}$: `INTR` → `t` → `1-s` (stands for 1 sample) → enter values of statistics → `Exe`
- In the "Data" row you usually choose option "Variable".
- If you are given sample observations choose "List" in the "Data" row and indicate the list number where you data are put.
- $\bar{X}_1 - \bar{X}_2 \pm z_{\frac{\alpha}{2}} \cdot \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$: `INTR` → `z` → `2-s` → enter values of statistics → `Exe`
- $\bar{X}_1 - \bar{X}_2 \pm t_{\frac{\alpha}{2}}(\text{df}) \cdot s_p \cdot \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$: `INTR` → `t` → `2-s` → enter values of statistics, `[Pooled: On]` → `Exe`
- $\bar{X}_1 - \bar{X}_2 \pm t_{\frac{\alpha}{2}}(\text{df}) \cdot \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$: `INTR` → `t` → `2-s` → enter values of statistics, `[Pooled: Off]` → `Exe`
- $\mu_{\text{diff}} = \bar{d} \pm t_{\frac{\alpha}{2}}(n-1) \cdot \frac{s_d}{\sqrt{n}}$ is calculated the same way the CI for μ with unknown σ (with t-statistic)
- $p = \hat{p} \pm z_{\frac{\alpha}{2}} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$: `INTR` → `z` → `1-p` → indicate `x` and `n` ($\hat{p} = \frac{x}{n}$) → `Exe`

- $p_1 - p_2 = \hat{p}_1 - \hat{p}_2 \pm z_{\frac{\alpha}{2}} \cdot \sqrt{\frac{\hat{p}_1 \cdot (1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2 \cdot (1-\hat{p}_2)}{n_2}}$: `INTR` → `z` → `2-p` → indicate $\mathtt{x}_1$, $\mathtt{n}_1$, $\mathtt{x}_2$, $\mathtt{n}_2$ → `Exe`

Full-score strategy

- Step 1. Introduce your variable.
- Step 2. Write down the given statistics of it.
- Step 3. Check/state necessary assumptions.
- Step 4. The formula. Find the value of z- or t-statistic.
- Step 5. Make calculations. Provide the answer.
- Step 6. Interpret your CI.

Top secret information

The simplest idea for estimating σ^2 is to sum the squared deviations of X_1 and X_2 and then estimate average deviation. Note that X_1 and X_2 have different means, so for each variable deviations are calculated with respect to different centers.

$$s_p^2 = \frac{\sum_{i=1}^{n_1} (x_{i1} - \bar{x}_1)^2 + \sum_{i=1}^{n_2} (x_{i2} - \bar{x}_2)^2}{n_p}$$

n_p is the overall number of observations in the two samples necessary to calculate the average squared deviation.

Since the standard deviation in each sample is calculated by the formula: $s^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}$ we can express the sums of squared deviations in terms of s_1 and s_2:

$$s_p^2 = \frac{s_1^2 \cdot (n_1 - 1) + s_2^2 \cdot (n_2 - 1)}{n_p}$$

Since s_1^2 and s_2^2 are unbiased estimators for σ^2 it is quite intuitive that $n_p = n_1 + n_2 - 2$ if we divide simply by $(n_1 + n_2)$, s_p^2 will underestimate σ^2.

Note that squared pooled standard deviation s_p^2 is also approximately equal to the average of s_1^2 and s_2^2 weighted by the corresponding sample sizes.

Sample AP practice problems

Problem 1. AP 2016 №5

A polling agency showed the following two statements to a random sample of 1,048 adults in the United States.

- *Environment statement*: Protection of the environment should be given priority over economic growth.
- *Economy statement*: Economic growth should be given priority over protection of the environment.

The order in which the statements were shown was randomly selected for each person in the sample. After reading the statements, each person was asked to choose the statement that was most consistent with his or her opinion. The results are shown in the table.

	Environment statement	Economy statement	No preference
Percent of sample	58%	37%	5%

(a) Assume the conditions for inference have been met. Construct and interpret a 95 percent confidence interval for the proportion of all adults in the United States who would have chosen the *economy statement*.

(b) One of the conditions for inference that was met is the number who choose the economy statement and the number who did not choose the economy statement are both greater than 10. Explain why it is necessary to satisfy that condition.

(c) A suggestion was made to use a two-sample z-interval for a difference between proportions to investigate whether the difference in proportions between adults in the United States who would have chosen the environment statement and adults in the United States who would have chosen the economy statement is statistically significant. Is the two-sample z-interval for a difference between proportions an appropriate procedure to investigate the difference? Justify your answer.

Solution

(a) Step 1. Let p be the proportion of all adults in the United States who would have chosen the economy statement.
Step 2. $\hat{p} = 0.37$, n=1048.
Step 3. Conditions which are said to be met, are that the sample was obtained by SRS and that it was large enough to ensure that $\hat{p} \sim N$ approximately (by CLT).
Step 4. $p = \hat{p} \pm z_{\frac{\alpha}{2}} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$, $z_{\frac{\alpha}{2}} \approx 1.96$
Step 5. $p = 0.37 \pm 1.96 \sqrt{\frac{0.37(1-0.37)}{1048}}$. $p \in [0.340; 0.400]$ at 95% confidence level.
Step 6. Thus, we are 95% sure that the population proportion of all adults in

the United States who would have chosen the economy statement is between 0.34 and 0.4.

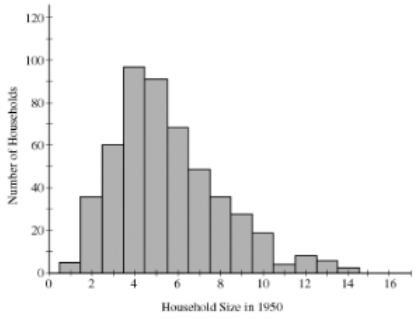
(b) The formula we use is based on the statement that $\hat{p} \sim N\left(p; \sqrt{\frac{p(1-p)}{n}}\right)$. The requirements $n\hat{p} > 10$ and $n(1-\hat{p}) > 10$ are established **to check whether the sample is large enough** to apply CLT which ensures that $\hat{p} \sim N$ approximately. So, we need to check the requirement to be sure that the mechanics are applicable.

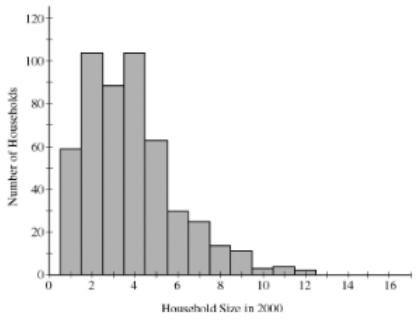
(c) The two-sample z-interval for a difference in proportions **is not appropriate** here. One of the conditions that should be satisfied in order to apply this procedure is independence of samples. In the given example both proportions are taken from the same sample, therefore they are not independent.

The procedure is derived from the normal distribution of difference of the two sample proportions. The difference of two normal random variables (both proportions are normal by CLT) would be normal given that they are independent. If this is not true, we cannot apply the usual formula.

Problem 2. AP 2012 №3

Independent random samples of 500 households were taken from a large metropolitan area in the United States for the years 1950 and 2000. Histograms of household size (number of people in a household) for the years are shown below.





(a) A researcher wants to use these data to construct a confidence interval to estimate the change in mean household size in the metropolitan area from the year 1950 to the year 2000. State the conditions for using a two-sample t-procedure, and explain whether the conditions for inference are met.

Solution

(a) To estimate the confidence interval for difference in mean household size in the metropolitan area from the year 1950 to the year 2000 we need to ensure the following conditions are satisfied:

1 Samples are SRS;

2 Samples and independent;

3 Samples are large enough or the variable of interest in normal.

The first condition is plausible: the samples are taken randomly and populations are most probably bigger than 5000. The second condition also seems to be satisfied since the two samples were taken in 1950 and 2000. Finally, third condition is satisfied because both samples are large enough ($500 > 30$) to state the sample means of household size are normal by CLT. The skewness of household sizes does not matter, as we only need sample means to be normal to apply the procedure.

Problem 3. AP 2011 Form B №5
During a flu vaccine shortage in the United States, it was believed that 45 percent of vaccine-eligible people received flu vaccine. The results of a survey given to a random sample of 2,350 vaccine-eligible people indicated that 978 of the 2,350 people had received flu vaccine.

(a) Construct a 99 percent confidence interval for the proportion of vaccine-eligible people who had received flu vaccine. Use your confidence interval to comment on the belief that 45 percent of the vaccine-eligible people had received flu vaccine.

(b) Suppose a similar survey will be given to vaccine-eligible people in Canada by Canadian health officials.

(c) A 99 percent confidence interval for the proportion of people who will have received flu vaccine is to be constructed. What is the smallest sample size that can be used to guarantee that the margin of error will be less than or equal to 0.02?

Solution

(a) Step 1. Let p be the proportion of vaccine-eligible people who had received flu vaccine.
Step 2. $\hat{p} = \frac{978}{2350} \approx 0.416$, $n = 2350$.
Step 3. We assume that the sample was obtained by SRS. Since $n\hat{p} = 978 > 10$ and $n(1-\hat{p}) = 1372 > 10$, the sample is large enough to ensure that $\hat{p} \sim N$ approximately (by CLT).
Step 4. $p = \hat{p} \pm z_{\frac{\alpha}{2}}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$, $z_{\frac{\alpha}{2}} \approx 2.576$
Step 5. $p = 0.416 \pm 2.576\sqrt{\frac{0.416(1-0.416)}{2350}}$. $p \in [\mathbf{0.390; 0.442}]$ at 99% confidence

level.
Step 6. Thus, we are 99% sure that the population proportion of vaccine-eligible people who had received flu vaccine is between 0.390 and 0.442. Since 0.45 does not belong to the confidence interval, **the evidence does not support the belief** that 45 percent of the vaccine-eligible people had received flu vaccine.

(b) Let p_c be the population proportion of vaccine-eligible people in Canada who had received flu vaccine. Margin of error $= z_{\frac{\alpha}{2}}\sqrt{\frac{\hat{p}_c(1-\hat{p}_c)}{n_c}} = 2.576\sqrt{\frac{\hat{p}_c(1-\hat{p}_c)}{n}}_c \leq 0.02$. $n_c \geqslant \left(\frac{2.576\cdot\sqrt{\hat{p}_c(1-\hat{p}_c)}}{0.02}\right)^2$. The minimum required n_c will depend on $\hat{p}_c$. When $\hat{p}_c\,(1-\hat{p}_c)$ takes its maximum value, the requirements for n_c are the highest. If we take them, n_c would be enough for any other situation to guarantee that Margin of error $\leqslant 0.02$. $\max\{\hat{p}_c\,(1-\hat{p}_c)\}$ is achieved when $\hat{p}_c = 0.5$. Thus, $n_c \geqslant \left(\frac{2.576\cdot\sqrt{0.5(1-0.5)}}{0.02}\right)^2 = 4,147.36$ will guarantee that margin of error does not exceed 0.02 for any $\hat{p}_c$. **Sample size whould be at least 4,148 people.**

Problem 4. AP 2010 №3
A humane society wanted to estimate with 95 percent confidence the proportion of households in its county that own at least one dog.

(a) Interpret the 95 percent confidence level in this context.

The humane society selected a random sample of households in its county and used the sample to estimate the proportion of all households that own at least one dog. The conditions for calculating a 95 percent confidence interval for the proportion of households in this county that own at least one dog were checked and verified, and the resulting confidence interval was 0.417 ± 0.119.

(b) A national pet products association claimed that 39 percent of all American households owned at least one dog. Does the humane society's interval estimate provide evidence that the proportion of dog owners in its county is different from the claimed national proportion? Explain.

(c) How many households were selected in the humane society's sample? Show how you obtained your answer.

Solution

(a) The 95 percent confidence level means that if one were to repeatedly take random samples of the same size from the population and construct a 95 percent confidence interval from each sample, then in the long run **95 percent of those intervals** would succeed in **capturing** the **actual** value of the population proportion of households in the county that own at least one dog.

(b) No. The 95 percent confidence interval 0.417 ± 0.119 is the interval $(0.298, 0.536)$. This interval includes the value 0.39 as a plausible value for the population proportion of households in the county that own at least one dog. Therefore, the confidence interval **does not provide evidence** that the **proportion** of dog owners in this county **is different** from the claimed national proportion.

(c) The sample proportion is 0.417, and the margin of error is 0.119. Determining the sample size requires solving the equation

$$0.119 = 1.96\sqrt{\frac{0.417\ (1-\ 0.417)}{n}}$$

$$0.119 = 1.96\frac{0.493}{\sqrt{n}}$$

$$0.119 = \frac{0.966}{\sqrt{n}}$$

$$\sqrt{n} = 0.966/0.119 = 8.12$$

$$n = \mathbf{65.95}$$

So the humane society must have selected **66** households for its sample.

Problem 5. AP 2009 Form B №6

Two treatments, A and B, showed promise for treating a potentially fatal disease. A randomized experiment was conducted to determine whether there is a significant difference in the survival rate between patients who receive treatment A and those who receive treatment B. Of 154 patients who received treatment A, 38 survived for at least 15 years, whereas 16 of the 164 patients who received treatment B survived at least 15 years. A can be administered only as a pill, and treatment B can be administered only as an injection.

(b) The conditions for inference have been met. Construct and interpret a 95 percent confidence interval for the difference between the proportion of the population who would survive at least 15 years if given treatment A and the proportion of the population who would survive at least 15 years if given treatment B.

In many of these types of studies, physicians are interested in the ratio of survival probabilities, $\frac{p_A}{p_B}$, where p_A represents the true 15-year survival rate for all patients who receive treatment A and p_B represents the true 15-year survival rate for all patients who receive treatment B. This ratio is usually referred to as the relative risk of the two treatments.

For example, a relative risk of 1 indicates the survival rates for patients receiving the two treatments are equal, whereas a relative risk of 1.5 indicates that the survival rate for patients receiving treatment A is 50 percent higher than the survival rate for patients receiving treatment B. An estimator of the relative risk is the ratio of estimated probabilities, $\frac{\hat{p}_A}{\hat{p}_B}$.

(c) Using the data from the randomized experiment described above, compute the estimate of the relative risk.

The sampling distribution of $\frac{\hat{p}_A}{\hat{p}_B}$ is skewed. However, when both sample sizes n_A and n_B are relatively large, the distribution of $\ln\left(\frac{\hat{p}_A}{\hat{p}_B}\right)$ — the natural logarithm

of relative risk — is approximately normal with a mean of $\ln\left(\frac{p_A}{p_B}\right)$ and a standard deviation of $\sqrt{\frac{1-p_A}{n_A p_A} + \frac{1-p_B}{n_B p_B}}$, where p_A and p_B can be estimated by using $\hat{p}_A$ and $\hat{p}_B$.

When a 95 percent confidence interval $\ln\left(\frac{p_A}{p_B}\right)$ is known, an approximate 95 percent confidence interval for $\frac{p_A}{p_B}$ — the relative risk of the two treatments — can be constructed by applying the inverse of the natural logarithm to the endpoints of the confidence interval for $\ln\left(\frac{p_A}{p_B}\right)$.

(d) The conditions for inference are met for the data in the experiment above, and a 95 percent confidence interval for $\ln\left(\frac{p_A}{p_B}\right)$ is $(0.3868, 1.4690)$. Construct and interpret a 95 percent confidence interval for the relative risk, $\frac{p_A}{p_B}$, of the two treatments.

(e) What is an advantage of using the interval in part (d) over using the interval in part (b)?

Solution

(b) Step 1. Let p_A and p_B represent the true 15-year survival rates for all patients who receive treatment A and treatment B, correspondingly.

Step 2. $\hat{p}_A = \frac{38}{154} \approx 0.247$, $n_A = 154$, $\hat{p}_B = \frac{16}{164} \approx 0.098$, $n_A = 164$.

Step 3. All the conditions are said to be met: both samples are SRS. Samples are independent. Samples are large enough ($n\hat{p} > 10$ and $n(1-\hat{p}) > 10$ for both treatments).

Step 4. $p_A - p_B = \hat{p}_A - \hat{p}_B \pm z_{\frac{\alpha}{2}} \cdot \sqrt{\frac{\hat{p}_A \cdot (1-\hat{p}_A)}{n_A} + \frac{\hat{p}_B \cdot (1-\hat{p}_B)}{n_B}}$, $z_{\frac{\alpha}{2}} = z_{0.025} \approx 1,96$.

Step 5. $p_A - p_B = 0.247 - 0.098 \pm 1.96 \cdot \sqrt{\frac{0.247 \cdot (1-0.247)}{154} + \frac{0.098 \cdot (1-0.098)}{164}}$, $p_A - p_B \in$ $(\mathbf{0.067}, \mathbf{0.231})$ at 95% confidence level.

Step 6. We are 95% sure that the difference between the proportion of the population who would survive at least 15 years if given treatment A and the corresponding proportion for treatment B is between 0.067 and 0.231. Since the whole interval is above zero, this suggests that treatment A is associated with higher survival rate.

(c) The estimate of relative risk is $\frac{\hat{p_A}}{\hat{p_B}} = \frac{0.247}{0.098} \approx \mathbf{2.529}$.

(d) Since $\ln\left(\frac{p_A}{p_B}\right) \in (0.3868, 1.4690)$, the corresponding boundaries of the confidence interval for $\frac{p_A}{p_B}$ are $e^{0.3868}$ and $e^{1.4690}$. Thus, $\frac{p_A}{p_B} \in \mathbf{(1.472, 4.345)}$ at 95% confidence level. People are between 1.47 and 4.34 times more likely to survive 15 or more years with treatment A than with treatment B.

(e) When the proportions of people who survive are low, as is the case with 0.25 and 0.10, **it may be more meaningful** or vivid to know that a patient's chance of survival with treatment A is 1.47 to 4.34 times what it would be with treatment B rather than to know that the difference in the proportions of people who survive is 0.07 to 0.23, which does not sound like very much.

Problem 6. AP 2008 Form B №3
A car manufacturer is interested in conducting a study to estimate the mean stopping distance for a new type of brakes when used in a car that is traveling at 60 miles per hour. These new brakes will be installed on cars of the same model and the stopping distance will be observed. The cost of each observation is \$100. A budget of \$12,000 is available to conduct the study and the goal is to carry it out in the most economical way possible. Preliminary studies indicate that $\sigma = 12$ feet for stopping distances.

(a) Are sufficient funds available to estimate the mean stopping distance to within 2 feet of the true mean stopping distance with 95% confidence? Explain your answer.

(b) A regulatory agency requires a 95% level of confidence for an estimate of mean stopping distance that is within 2 feet of the true mean stopping distance. The car manufacturer cannot exceed the budget of \$12,000 for the study. Discuss the consequences of these constraints.

Solution

(a) Let X be a stopping distance for a new type of brakes when used in a car that is traveling at 60 miles per hour. We know that σ=12. The appropriate confidence interval for the mean is: $\mu = \bar{X} \pm z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}$, $z_{\frac{\alpha}{2}} = z_{0.025} \approx 1.96$. There is requirement: margin of error$\leq$2, $z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}} \leq 2$.

$n \geqslant \left(\frac{z_{\frac{\alpha}{2}} \cdot \sigma}{2}\right)^2 = \left(\frac{1.96 \cdot 12}{2}\right)^2 \approx 138.298.$

Thus, the experiment should be repeated at least 139 times.

The budget allows to conduct $n = \frac{12{,}000}{100} = 120$ experiments, which is less than 139. Therefore, **there no sufficient funds.**

(b) To be within 2 feet of the true mean with 95% confidence, 139 observations are required. The budget of \$12,000 only allows 120 observations to be taken. Therefore, **the company will not be able to meet the regulatory agency's requirements** with the allocated budget.

Problem 7. AP 2005 №6
Lead, found in some plants, is a neurotoxin that can be especially harmful to the developing brain and nervous system of children. Children frequently put their hands in their mouth after touching painted surfaces, and this is the most common type of exposure to lead.

A study was conducted to investigate whether there were differences in children's exposure to lead between suburban day-care centers and urban day-care centers in one large city. For this study, researchers used a random sample of 20 children in suburban day-care centers. Ten of these 20 children were randomly selected to play outside; the remaining 10 children played inside. All children had their hands wiped clean before beginning their assigned one-hour play period either outside or inside. After the play period ended, the amount if lead in micrograms (mcg) on each child's dominant hand was recorded.

The mean amount of lead on the dominant hand for the children playing inside was 3.75 mcg, and the mean amount of lead for the children playing inside was 5.65 mcg. A 95 percent confidence interval for the difference in the mean amount of lead after one hour inside versus one hour outside was calculated to be (-2.46, -1.34).

A random sample of 18 children in urban day-care centers in the same large city was selected. For this sample, the same process was used, including randomly assigning children to play inside or outside. The data for the amount (in mcg) of lead in each child's dominant hand are shown in the table below.

Urban day-care centers

Inside	6	5	4	4	4.5	5	4.5	3	5
Outside	15	25	18	14	20	13	11	22	20

(d) Use a 95 percent confidence interval to estimate the difference in the mean amount of lead on child's dominant hand after one hour of play inside versus an hour of play outside at urban day-care centers in this city. Be sure to interpret your interval.

Solution

(d) Step 1. Let X_1 and X_2 be amounts of lead on child's dominant hand after one hour of play inside and outside, correspondingly at urban day-care centers in this city.

Step 2. We calculated that: $\bar{X}_1 \approx 4.556$, $\bar{X}_2 \approx 17.556$, $s_1 = 0.846$, $s_2 = 4.613$, $n_1 = n_2 = 9$.

Step 3. 1). It's reasonable to assume that populations of children playing inside and outside at urban day-care centers in this city are large enough (at least 90), so that the samples are no more than 10% of the populations. We also assume both samples were obtained by SRS. 2). The samples are not large enough ($n_1 = n_2 = 9 \leqslant 30$) to ensure that $\bar{X}_1$ and $\bar{X}_2$ are normal by CLT. Therefore we need to check that amounts of lead X_1 and X_2 can themselves be approximated by normal distribution. Boxplots of both datasets are rather symmetric. There are not outliers on both datasets. So, assumption of normality is not unreasonable. 3). Since children were randomly divided into two groups the two samples are independent.

[Draw a boxplot here]

Step 4. $\mu_1 - \mu_2 = \bar{X}_1 - \bar{X}_2 \pm t_{\frac{\alpha}{2}}(\text{df}) \cdot \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$, $\text{df} \approx 8.537$. $t_{\frac{\alpha}{2}}(\text{df}) \approx t_{0.025}(8.5) \approx 2.281$

Step 5. $\mu_1 - \mu_2 = 4.556 - 17.556 \pm 2.281 \cdot \sqrt{\frac{0.846^2}{9} + \frac{4.613^2}{9}}$. $\boldsymbol{\mu_1 - \mu_2} \in$ $[\mathbf{-16.566; -9.434}]$ at 95% confidence level.

Step 6. We are 95% sure that the difference in mean times devoted to homework by sixth- and seventh-grade students is between -16.566 and -9.434. Since the whole interval lies below zero, we can conclude that children playing outside are more exposed to lead influence.

Problem 8. AP 2004 №6

A pharmaceutical company has developed a new drug to reduce cholesterol. A regulatory agency will recommend the new drug for use if there is convincing evidence that the mean reduction in cholesterol level after one month of use is more than 20 milligrams/deciliter (mg/dl), because a mean reduction of this magnitude would be greater than the mean reduction for the current most widely used drug.

The pharmaceutical company collected data by giving the new drug to a random sample of 50 people from the population of people with high cholesterol. The reduction in cholesterol level after one month of use was recorded for each individual in the sample, resulting in a sample mean reduction and standard deviation of 24 mg/dl and 15 mg/dl, respectively.

(a) The regulatory agency decides to use an interval estimate for the population mean reduction in cholesterol level for the new drug. Provide this 95 percent confidence interval. Be sure to interpret this interval.

(c) The company would like to determine a value L that would allow them to make the following statement.

We are 95 percent confident that the true mean reduction in cholesterol level is greater than L.

A statement of this form is called a one-sided confidence interval. The value of L can be found using the following formula.

$$L = \bar{X} - t^* \frac{s}{\sqrt{n}}$$

This has the same form as the lowest endpoint of the confidence interval in part (a), but requires a different critical value, t^*. What value should be used for t^*?

Recall that the sample mean reduction in cholesterol level and standard deviation are 24 mg/dl and 15 mg/dl, respectively. Compute the value of L.

(d) If the regulatory agency had used the one-sided confidence interval in part (c) rather than the interval constructed in part (a), would it have reached a different conclusion? Explain.

Solution

(a) Step 1. Let X be a reduction of cholesterol level.

Step 2. $\bar{X} = 24$, $s = 15$, $n = 50$.

Step 3. Let's assume that the sample was obtained by SRS. Since $n > 30$, the sample is large enough to state that sample would be approximately normal.

Step 4. $\mu = \bar{X} \pm t_{\frac{\alpha}{2}}(n-1) \cdot \frac{s}{\sqrt{n}}$, $t_{\frac{\alpha}{2}}(n-1) = t_{0.025}(49) \approx 2.010$.

Step 5. $\mu = 24 \ \pm 2.01 \cdot \frac{15}{\sqrt{50}}$. Thus, $\mu \in [\mathbf{19.737; 28.263}]$ at 95% confidence level.

Step 6. Thus, we are 95% sure that the true mean reduction in cholesterol level is between 19.737 mg/dl and 28.263 mg/dl. Since the interval includes 20, there is no convincing evidence that the mean reduction in cholesterol level after one month of use is more than 20 mg/dl.

(c) The critical value t^* should ensure that above are the 95% of highest observations and below are 5% of lowest observations. Thus, $t^* = t_{0.05}(49) \approx \mathbf{1.677}$.

Thus, $L = \bar{X} - t^* \frac{s}{\sqrt{n}} \approx 24 - 1.677 \cdot \frac{15}{\sqrt{50}} \approx \mathbf{20.443}$.

(d) Since the one-sided CI has shown that $\mu \geqslant 20.443$, we can conclude that here is convincing evidence that the mean reduction in cholesterol level is more than 20 mg/dl. Thus, **when using the answer in (c) the agency would have reached different conclusion than in (a)**.

Problem 9. AP 2004 Form B №4
The principal at Crest Middle School, which enrolls only sixth-grade students and seventh-grade students is interested in determining how much time students at that school spend on homework each night. The table below shows the mean and standard deviation of the amount of time spent on homework each night (in minutes) for a random sample of 20 sixth-grade students and a separate random sample of 20 seventh-grade students at this school.

	Mean	Standard deviation
Sixth-grade students	27.3	10.8
Seventh-grade students	47.0	12.4

Based on dotplots of these data it is not unreasonable to assume that the distribution of times for each grade were approximately normally distributed.

(a) Estimate the difference in mean times spent on homework for all sixth- and seventh-grade students in this school using an interval. Be sure to interpret your interval.

(b) An assistant principal reasoned that a much narrower confidence interval could be obtained if the students were paired based on their responses; for example, pairing

the sixth-grade student and the seventh-grade student with the highest number of minutes spent on homework, and so on. Is the assistant principal correct in thinking that matching students in this way and then computing a matched-pairs confidence interval for the mean difference in time spent on homework is a better procedure than the one used in part (a)? Explain why or why not.

Solution

(a) Step 1. Based on dotplots of these data it is not unreasonable to assume that the distribution of times for each grade were approximately normally distributed.

Step 2. $\bar{X}_1 = 27.3$, $\bar{X}_2 = 47.9$, $s_1 = 10.8, s_2 = 12.4, n_1 = n_2 = 20$.

Step 3. 1). We assume that populations of both sixth- and seventh-grade students in this school are at least 200, so that the sample of size 20 is no more than 10% of the population. We also assume both samples were obtained by SRS. 2). The samples are not large enough ($n_1 = n_2 = 20 < 30$) to ensure that $\bar{X}_1$ and $\bar{X}_2$ are normal by CLT. Therefore we need to assume that times spent on homework (X_1 and X_2) are themselves normally distributed. The shape of distribution advocates that this assumption is not unreasonable. 3). Let's assume the two samples are independent.

Step 4. $\mu_1 - \mu_2 = \bar{X}_1 - \bar{X}_2 \pm t_{\frac{\alpha}{2}}(df) \cdot \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$, df ≈ 37.297. Let's choose 95% confidence level. Then, $t_{\frac{\alpha}{2}}$ (df) $\approx t_{0.025}\,(37) \approx 2.026$.

Step 5. $\mu_1 - \mu_2 = 27.3 - 47 \pm 2.026 \cdot \sqrt{\frac{10.8^2}{20} + \frac{12.4^2}{20}}$. $\boldsymbol{\mu_1 - \mu_2 \in [-27.148; -12.252]}$ at 95% confidence level.

Step 6. We are 95% sure that the difference in mean times devoted to homework by sixth- and seventh-grade students is between -28.048 and -13.152.

Since the whole interval lies below zero, there is statistical evidence that the difference in population proportions is negative which means seventh-grade students tend to spend more time on homework.

(b) This is not an appropriate procedure, assistant principal is **not correct**. We are interested in difference in study times by sixth- and seventh-grade students. If we pair them according to their times, we'll see smaller difference, which would underestimate the true value. The appropriate way to pair students might be done by matching students with similar scores, family income, distance between home and the school, interests etc, or other factors related to the response, but not response itself. Even a better strategy would be to take two datasets from the same students as they study in 6 and 7 grade and calculate individual differences in times.

Problem 10. AP 2003 №6

The Blue Shell Shuttle Bus Company has recently acquired the rights to run a shuttle between Lonestar's hotels and it's airport, which is several miles away. For the new

route, the company has a choice of running coaches that can carry up to 60 people or smaller ones that can carry up to 12 people. The company has a policy that each of its routes is served only by one type of shuttle vehicle. In addition, due to the allocation of their vehicles, to other routes, no change in their decision can be considered for at least a year. The annual return (profit or loss) depends on whether the demand for the service is strong or weak. Research suggests that the following returns can be expected.

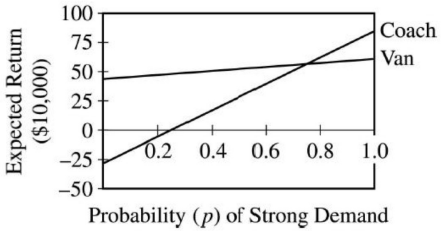
Annual return ($10,000)		
Vehicle decision	Demand	
	Strong	Weak
Coach	84	-27
Van	61	45

For instance, if a coach is used and demand is strong, the expected annual return is $840,000. The expected return to the company can be calculated based on the probability of the string demand. Let p represent the probability of strong demand; then $(1-p)$ represents the probability of weak demand.

An equation that can be used to compute the expected return from the use of coaches based on the value of p is

$$84p + (-27)(1-p) = 111p - 27$$

An equation that can be used to compute the expected return from the use of vans based on the value of p is $61p + 45(1-p) = 16p + 45$ These two functions are shown on the graph below.



(a) The value of p for which the expected annual return for the vans is equal to the expected annual return for the coaches is 0.76. If the probability of strong demand is less than this value, which decision, running coaches or running vans, will provide the greater expected return? Justify your answer.

(b) There are several thousand markets similar to Lonestar's market across the country. A random sample of 100 of these markets reveals that the demand for airport shuttle is strong in 65% of them and the demand in the remaining 35 is weak. Using the results of this sample, construct and interpret a 95 percent confidence interval for the proportion of similar markets that will experience a strong demand.

(c) The president of the Blue Shell has decided to use vans for the new route. Using the results of the analysis in parts (a) and (b), write a few sentences to justify this decision.

(d) After looking at the interval in part (b) and considering possible annual returns, the vice president of the Blue Shell believes that the president has made an incorrect decision in choosing to use vans. Explain has this conflicting position could be supported.

Solution

(a) As it can be seen from the graph, when p is less than 0.76 expected return of vans is higher than the expected return of coaches (the line denoting return for vans is higher). So, choosing vans will provide greater return on average.

(b) Step 1 is not needed here as p is introduced.

Step 2. $n = 100$, $\hat{p} = 0.65$.

Step 3. Let's assume the sample was obtained by SRS. We see that the sample (100) is less than 10% of the population (several thousands). Also, since $n\hat{p} = 65 > 10$ and $n(1-\hat{p}) = 35 > 10$ the sample is large enough to ensure that $\hat{p} \sim N$ approximately (by CLT).

Step 4. $p = \hat{p} \pm z_{\frac{\alpha}{2}}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$, $z_{\frac{\alpha}{2}} \approx 1.96$

Step 5. $p = 0.65 \pm 1.96\sqrt{\frac{0.65(1-0.65)}{100}}$. $p \in [0.557; 0.743]$ at 95% confidence level.

Step 6. Thus, we are 95% sure that the true proportion of markets where the demand is strong is between 0.557 and 0.743.

(c) All values in the confidence interval are below 0.76. Thus, most probability $p < 0.76$. As we've explained in (a) in this case expected return is higher when using vans. That's why this is a preferable decision.

(d) The demand in the Lonestar's market will either be strong or weak. Since the confidence interval contains only values above 0.5, it is more likely to be strong. If demand is strong, using coaches will provide \$810,000 of return, while using vans will only provide \$610,000 (based on the data in the table. Since the demand is more likely to be strong than weak and coaches produce higher returns in conditions of strong demand, choosing coaches is a better decision.

Problem 11. AP 2003 Form B №6

Researches at a large health maintenance organization (HMO) are planning a study of a certain mild illness. They will select a random sample of patients who are ages 35 to 54 and see if they contract the illness in the next year. The researchers are interested in estimating the proportions of men and of women who are likely to develop the illness in each of 4 age groups: 35-39, 40-44, 45-49, and 50-54.

The researches plan to include 2,000 patients in the study. Suppose the researchers draw a random sample from all of the patients at this HMO who are ages 35 to 54 and find the following numbers within each gender and age-group.

	Age-group			
	35-39	40-44	45-49	50-54
Male	350	230	150	60
Female	445	370	245	150

(a) Suppose that at the end of the study, 10 percent of the females in 40-44 age-group contracted the illness. Calculate a 95 percent confidence interval to estimate the population proportion of females in this age-group that contracted the illness.

Interpret this confidence interval in the context of this situation.

Interpret the confidence level of 95 percent.

(b) Suppose that at the end of the study, 10 percent of the males in 40-44 age-group contracted the illness. The corresponding 95 percent confidence interval to estimate the population proportion of males in this age-group that contracted the illness is (0.061; 0.139).

Note that this interval and the interval in part (a) are of different lengths even though the two sample proportions were identical. What would be an alternative way to allocate a sample of 2,000 subjects so that the 95 percent confidence interval widths for all male age-groups and for all female age-groups (i.e., for all 8 groups) would be the same when the sample proportions are the same? Justify your answer.

(c) Based on previous studies, researchers believe that the percentages of those who contract the illness will be similar for males and females, and therefore plan to ignore gender when selecting a sample for this study. Previous studies also indicate that the percentages of adults who will contract this illness in the 35-39, 40-44, 45-49, and 50-54 age-groups are anticipated to be 5%, 8%, 20%, and 35%, respectively. How should the sample of 2,000 subjects be allocated with respect to age-groups so that the widths of 95 percent confidence intervals for the four groups will be approximately the same? Justify your answer.

Solution

1. Step 1. Let p be the proportion of females in 40-44 age-group that contracted the illness.

   Step 2. $\hat{p} = 0.1$, $n = 370$

   Step 3. Assume that sample is SRS of all females in this age-group. Also, since $n\hat{p} = 37 > 10$ and $n(1-\hat{p}) = 333 > 10$, the sample is large enough to ensure that $\hat{p} \sim N$ approximately (by CLT).

   Step 4. $p = \hat{p} \pm z_{\frac{\alpha}{2}} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$, $z_{\frac{\alpha}{2}} \approx 1.96$.

Step 5. $p = 0.1 \pm 1.96\sqrt{\frac{0.1(1-0.1)}{370}}$. $p \in$ [**0.069; 0.131**] at 95% confidence level.

Step 6. Thus, we are 95% sure that the population proportion of females in 40-44 age-group that contract the illness is between 0.069 and 0.131.

The confidence level of 95 % indicates that if we were to take many simple random samples of size 370 and calculated a sample proportion $\hat{p}$ and the corresponding 95% confidence intervals based on each sample, 95% of such intervals would include the true proportion p.

2. To ensure that the widths of the intervals are the same we should make margin of error equal. Margin of error $= z_{\frac{\alpha}{2}}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$. Since sample proportions and level of confidence are the same for males and females, difference in widths is explained by different value of sample size: $n_{\text{Female}} = 370$, $n_{\text{Male}} = 230$. So, to ensure equal widths of the confidence intervals we should have equal number of people of the 8 groups. $\frac{2{,}000}{8} =$ **250 people** should be in each group.

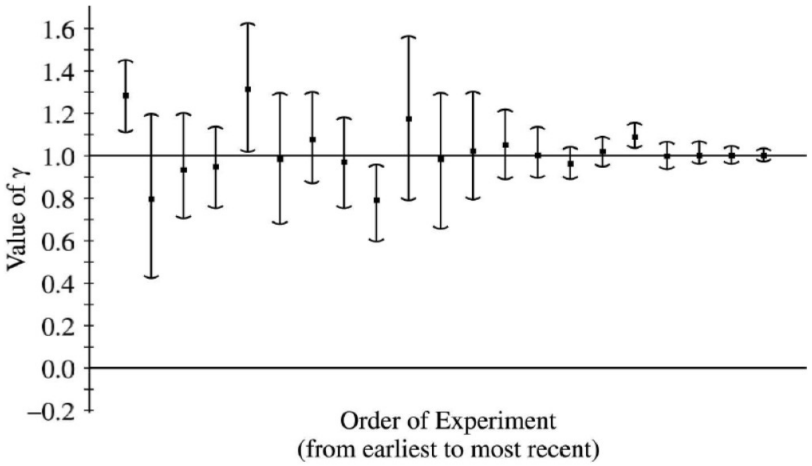
3. Let n_1, n_2, n_3 and n_4 be sample sizes for the four age categories that would ensure equal widths of confidence intervals. For that purpose we need the equality to hold: $\frac{0.05\cdot0.95}{n_1} = \frac{0.08\cdot0.92}{n_2} = \frac{0.2\cdot0.8}{n_3} = \frac{0.35\cdot0.65}{n_4}$. We also know that $n_1 + n_2 + n_3 + n_4 = 2,000$. Solving this system of equations we get sample sizes that provide aproximally equal margins of error: $n_1 = 187$, $n_2 = 289$, $n_3 = 629$, $n_4 = 895$.

Problem 12. AP 2002 №1

In 1915 Einstein's theory predicted that the curvature of space, denoted by γ, was 1, while Newtonian theory predicted it was 0. Since 1915 scientists have repeatedly found estimates of γ using various methods and procedures. Each estimate has a margin of error. The figure below displays

(estimate $\pm$ margin of error)

from each of 21 experiments.



(a) Based on the display, describe how the precision of the estimates of γ has changed over time.

(b) Write a few sentences describing the strength of evidence the experiments provide for the claim from Newtonian theory that $\gamma = 0$. Your response must include justification based on the display.

(c) Write a few sentences describing the strength of evidence the experiments provide for the claim from Einstein's theory that $\gamma = 1$. Your response must include justification based on the display.

Solution

1. We see that the intervals have become narrower over time. That means that margins of error became smaller. Thus, **estimates of γ have become more accurate**.

2. The picture shows that not a single interval contains 0. Thus, there is **strong evidence against** the claim that $\gamma = 0$.

3. The display shows that 17 intervals out of 21 do contain 1. Also, intervals seem to converge to 1 over time as they become more precise. Thus, there is **strong statistical evidence for the claim that $\gamma = 1$**.

Problem 13. AP 2000 №2

Anthropologists have discovered a prehistoric cave dwelling that contains a large number of adult human footprints. To study the size of the adults who used the cave dwelling, they randomly selected 20 of the footprints from the population of all footprints in the cave and measured the length of those footprints. Some statistics resulting from this random sample are as follows.

Sample size	20	Minimum	15.2 cm
Mean	24.8 cm	First Quartile	18.7 cm
Standard deviation	7.5 cm	Median	21.5 cm
		Third Quartile	30.0 cm
		Maximum	37.cm

The anthropologists would like to construct a 95 percent confidence interval for the mean foot length of the adults who used the cave dwelling.

(a) What assumptions are necessary in order for this confidence interval to be appropriate?

(b) Discuss whether each of the assumptions listed in your response to (a) appears to be satisfied in this situation.

Solution

1. First, we need to assume the sample is an SRS of human footprints from the population of adults living in a cave dwelling. Second, the sample is not large enough for CLT to guarantee that the sample mean of footprint length is approximately normal. Therefore, we need to assume the length is normally distributed.

2. The random sample of footprints was taken randomly from only one cave. This was also a random sample from population of footprints, not population of adults. Therefore, it may contain footprints of children or contain several footprints belonging to the same human. This means that observations are not independent, and the sample may contain observations from another population. Thus, **the first assumption is not satisfied.**

   The z-scores for the observed minimum, lower and upper quartiles and maximum:

   $z_{\min} = \frac{15.2-24.8}{7.5} = -1.28$, $z_{Q1} = \frac{18.7-24.8}{7.5} \approx -0.813$, $z_{Q2} = \frac{30-24.8}{7.5} \approx 0.693$, $z_{\max} = \frac{37-24.8}{7.5} \approx 1.627$.

   The corresponding z-scores for quartiles of the normal distribution are: $z_{Q1} \approx -0.675$, $z_{Q2} \approx 0.675$. We also know that for normal distribution 95% of observations lie between -1.96 and 1.96 z-scores.

   First, we see that z-scores for minimum and maximum and for first and third quartiles are not symmetric. Second, comparing observed z-scores to the z-scores of corresponding statistics of the normal distribution shows they are very different. For example -1.28 is much higher than -1.96.

   Thus, we come to the conclusion that **the observed data does not seem to come from normal distribution.**

Practice AP problems

Problem 1. AP 2017 №2
The manager of a local fast-food restaurant is concerned about customers who ask for a water cup when placing an order but fill the cup with a soft drink from the beverage fountain instead of filling the cup with water. The manager selected a random sample of 80 customers who asked for a water cup when placing an order and found that 23 of those customers filled the cup with a soft drink from the beverage fountain.

(a) Construct and interpret a 95 percent confidence interval for the proportion of all customers who, having asked for a water cup when placing an order, will fill the cup with a soft drink from the beverage fountain.

(b) The manager estimates that each customer who asks for a water cup but fills it with a soft drink costs the restaurant \$0.25. Suppose that in the month of June 3,000 customers ask for a water cup when placing an order. Use the confidence interval constructed in part (a) to give an interval estimate for the cost to the restaurant for the month of June from the customers who ask for a water cup but fill the cup with a soft drink.

Problem 2. AP 2015 №2
To increase business, the owner of a restaurant is running a promotion in which a customer's bill can be randomly selected to receive a discount. When a customer's bill is printed, a program in the cash register randomly determines whether the customer will receive a discount on the bill. The program was written to generate a discount with a probability of 0.2, that is, giving 20 percent of the bills a discount in the long run.

However, the owner is concerned that the program has a mistake that results in the program not generating the intended long-run proportion of 0.2.

The owner selected a random sample of bills and found that only 15 percent of them received discounts. A confidence interval for p, the proportion of bills that will receive a discount in the long run, is 0.15 ± 0.06.

All conditions for inference were met.

(a) Consider the confidence interval 0.15 ± 0.06

  (i) Does the confidence interval provide convincing statistical evidence that the program is not working as intended? Justify your answer.

  (ii) Does the confidence interval provide convincing statistical evidence that the program generates the discount with a probability of 0.2? Justify your answer.

A second random sample of bills was taken that was four times the size of the original sample. In the second sample 15 percent of the bills received the discount.

(b) Determine the value of the margin of error based on the second sample of bills that would be used to compute an interval for p with the same confidence level as that of the original interval.

(c) Based on the margin of error in part (b) that was obtained from the second sample, what do you conclude about whether the program is working as intended? Justify your answer.

Problem 3. AP 2013 №1b

An environmental group conducted a study to determine whether crows in a certain region were ingesting food containing unhealthy levels of lead. A biologist classified lead levels greater than 6.0 parts per million (ppm) as unhealthy. The lead levels of a random sample of 23 crows in the region were measured and recorded. The data are shown in the stemplot below.

Lead Levels

2	8
3	0
3	5 8 8
4	1 1 2
4	6 8 8
5	0 1 2 2 3 4
5	9 9
6	3 4
6	6 8

Key: 2|8 = 2.8 ppm

(a) What proportion of crows in the sample had lead levels that are classified by the biologist as unhealthy?

(b) The mean lead level of the 23 crows in the sample was 4.90 ppm and the standard deviation was 1.12 ppm. Construct and interpret a 95 percent confidence interval for the mean lead level of crows in the region.

Problem 4. AP 2011 №6

Every year, each student in a nationally representative sample is given tests in various subjects. Recently, a random sample of 9,600 twelfth-grade students from the United States were administered a multiple choice United States history exam. One of the multiple-choice questions is below. (The correct answer is C.)

In 1935 and 1936 the Supreme Court declared that important parts of the New Deal were unconstitutional. President Roosevelt responded by threatening to:

(A) impeach several Supreme Court justices

(B) eliminate the Supreme Court

(C) appoint additional Supreme Court justices who shared his views

(D) override the Supreme Court's decisions by gaining three-fourths majorities

in both houses of Congress

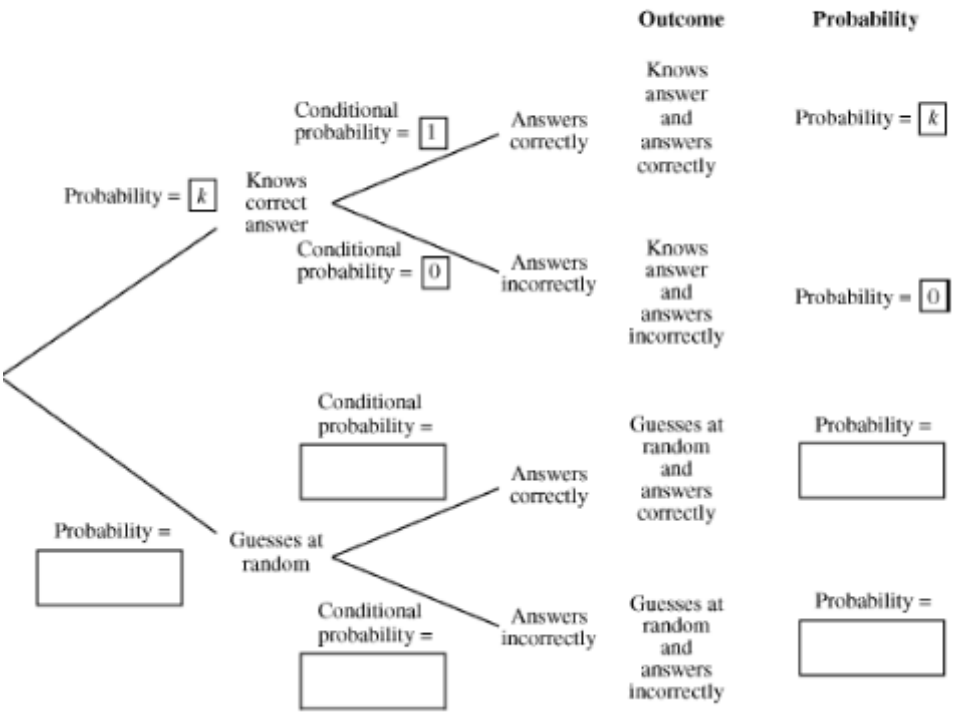
Of the 9,600 students, 28 percent answered the multiple-choice question correctly.

(a) Let p be the proportion of all United States twelfth-grade students who would answer the question correctly. Construct and interpret a 99 percent confidence interval for p.

Assume that students who actually know the correct answer have a 100 percent chance of answering the question correctly, and students who do not know the correct answer to the question guess completely at random from among the four options.

Let k represent the proportion of all United States twelfth-grade students who actually know the correct answer to the question.

TREE DIAGRAM OF OUTCOMES FOR A RANDOMLY SELECTED TWELFTH-GRADE STUDENT



(b) See Chapter 1.

(c) Using your interval from part (a) and your answer to part (c), calculate and interpret a 99 percent confidence interval for k, the proportion of all United States twelfth-grade students who actually know the answer to the history question. You may assume that the conditions for inference for the confidence interval have been checked and verified.

Problem 5. AP 2010 Form B №4
A husband and wife, Mike and Lori, share a digital music player that has a feature that randomly selects which song to play. A total of 2,384 songs were loaded onto the player, some by Mike and the rest by Lori. Suppose that when the player was in the random-selection mode, 13 of the first 50 songs selected were songs loaded by Lori.

(a) Construct and interpret a 90 percent confidence interval for the proportion of songs on the player that were loaded by Lori.

Problem 6. AP 2009 №4
One of the two fire stations in a certain town responds to calls in the northern half of the town, and the other fire station responds to calls in the southern half of the town. One of the town council members believes that the two fire stations have different mean response times. Response time is measured by the difference between the time an emergency call comes into the fire station and the time the first fire truck arrives at the scene of the fire.

Data were collected to investigate whether the council member's belief is correct. A random sample of 50 calls selected from the northern fire station had a mean response time of 4.3 minutes with a standard deviation of 3.7 minutes. A random sample of 50 calls selected from the southern fire station had a mean response time of 5.3 minutes with a standard deviation of 3.2 minutes.

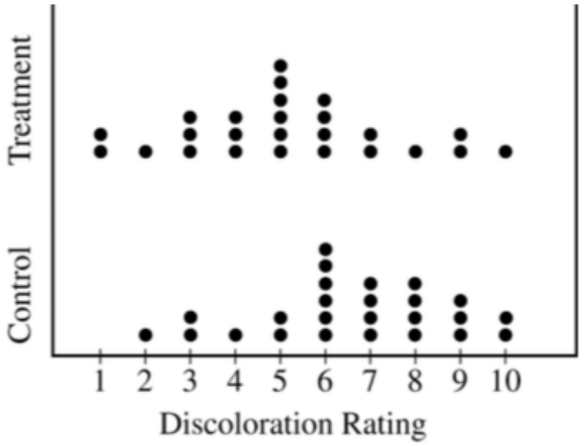
(a) Construct and interpret a 95 percent confidence interval for the difference in mean response times between the two fire stations.

(b) Does the confidence interval in part (a) support the council member's belief that the two fire stations have different mean response times? Explain.

Problem 7. AP 2007 №1
The department of agriculture at a university was interested in determining whether a preservative was effective in reducing discoloration in frozen strawberries. A sample of 50 ripe strawberries was prepared for freezing. Then the sample was randomly divided into two groups of 25 strawberries each. Each strawberry was placed into a small plastic bag.

The 25 bags in the control group were sealed. The preservative was added to the 25 bags containing strawberries in the treatment group, and then those bags were sealed. All bags were stored at 0°C for a period of 6 months. At the end of this time, after the strawberries were thawed, a technician rated each strawberry's discoloration from 1 to 10, with a low score indicating little discoloration.

The dotplots below show the distributions of discoloration rating for the control and treatment groups



1. The standard deviation of ratings for the control group is 2.141. Explain how this value summarizes variability in the control group

2. Based on the dotplots, comment on the effectiveness of the preservative in lowering the amount of discoloration in strawberries. (No calculations are necessary.)

3. Researchers at the university decided to calculate a 95 percent confidence interval for the difference in mean discoloration rating between strawberries that were not treated with preservative and those that were treated with preservative. The confidence interval they obtained was (0.16, 2.72). Assume that the conditions necessary for the t-confidence interval are met.

   Based on the confidence interval, comment on whether there would be a difference in the population mean discoloration ratings for the treated and untreated strawberries.

Problem 8. AP 2006 №4

Patients with heart-attack symptoms arrive at an emergency room either by ambulance or self-transportation provided by themselves, family, or friends. When a patient arrives at the emergency room, the time of arrival is recorded. The time when the patient's diagnostic treatment begins is also recorded.

An administrator of a large hospital wanted to determine whether the mean wait time (time between arrival and diagnostic treatment) for patients with heart-attack symptoms differs according to the mode of transportation. A random sample of 150 patients with heart-attack symptoms who had reported to the emergency room was selected. For each patient, the mode of transportation and wait time were recorded. Summary statistics for each mode of transportation are shown in the table below.

Mode of Transportation	Sample Size	Mean Wait Time (in minutes)	Standard Deviation of Wait Times (in minutes)
Ambulance	77	6.04	4.30
Self	73	8.30	5.16

(a) Use a 99 percent confidence interval to estimate the difference between the mean wait times for ambulance- transported patients and self-transported patients at this emergency room.

(b) Based only on this confidence interval, do you think the difference in the mean wait times is statistically significant? Justify your answer.

Problem 9. AP 2006 Form B №2
A large company has two shifts - a day shift and a night shift. Parts produced by the two shifts must meet the same specifications. The manager of the company believes that there is a difference in the proportion of parts produced with specifications by two shifts. To investigate this belief, random sample of parts that were produced on each of these shifts were selected. For the day shift, 188 of its 200 selected met the specifications.

(a) Use 96 percent confidence interval to estimate the difference in the proportions of parts produced within specifications by the two shifts.

(b) Based only in the confidence interval, do you think that the difference in the proportions of parts produced within specifications by two the shifts is significantly different from 0? Justify your answer.

Problem 10. AP 2005 №5b – Вбить условие

Problem 11. AP 2005 Form B №4
A researcher believes that treating seeds with certain additives before planting can enhance the growth of plants. An experiment to investigate this is conducted in a greenhouse. From a large number of Roma tomato seeds, 24 seeds are randomly chosen and 2 are assigned to each of 12 containers. One of the 2 seeds is randomly selected and treated with the additive. The other seed serves as a control. Both seeds are then planted in the same container. The growth, in centimeters, of each of the 24 plants is measured after 30 days. These data were used to generate the partial computer output shown below. Graphical displays indicate that the assumption of normality is not unreasonable.

	N	Mean	StDev	SE Mean
Control	12	15.989	1.098	0.317
Treatment	12	18.004	1.175	0.339
Difference	12	-2.015	1.163	0.336

(a) Construct a confidence interval for the mean difference in growth, in centimeters, of the plants from the untreated and treated seeds. Be sure to interpret this interval.

(b) Based only on the confidence interval in part (a), is there sufficient evidence to conclude that there is a significant mean difference in growth of the plants from untreated seeds and the plants from treated seeds? Justify your conclusion.

Problem 12. AP 2002, №6
A survey is given to a random sample of students at a university included a question about which of two well-known comedy shows, S or F, students preferred. The students were asked the question, "Do you prefer S or F?"

The responses are shown below.

Preference		
S	F	Total
185	139	324

(a) Based on the results of this survey, construct and interpret a 95% confidence interval for the proportion of students in the population who would respond S to the question, "Do you prefer S or F?"

(b) What is the meaning of "95% confidence" in part (a)?

Problem 13. AP 2002 Form B №4
Each person in a random sample of 1,026 adults in the United States was asked the following question.

"Based on what you know about the Social Security system today, what would you like Congress and the President to do during this next year?"

The response choices and the percentages selecting them are shown below.

Completely overhaul the system	19%
Make some major changes	39%
Make some minor adjustments	30%
Leave the system the way it is now	11%
No opinion	1%

(a) Find a 95% interval for the proportion of all United States adults who would respond "Make some major changes" to the question. Give an interpretation of the confidence level.

(b) An advocate for leaving the system as it is now commented, "Based on this poll, only 39% of adults in the sample responded that they want some major changes made to the system, while 41% responded that they want only minor changes or no changes at all. Therefore, we should not change the system." Explain why this statement, while technically correct, is misleading.

Answers to the practice problems

Problem 1. (a) (0.1883 , 0.3867), (b) (141.25$, 290.00$)

Problem 2. (a) (i) No (ii) No, (b) 0.03, (c) There is statistical evidence that the program is not working as intended and is not generating discounts with a probability of 0.2.

Problem 3. (a) 0.174, (b) (4.411 , 5.3803)

Problem 4. (a) (0.268 , 0,292), (d) (0.024 , 0.056)

Problem 5. (0.158 , 0.362)

Problem 6. (a) The standard deviation measures a typical distance between individual ratings and average rating for the strawberries in the control group.

Problem 7. The preservative is effective in lowering the amount of discoloration in strawberries.

Problem 8. (a) (0.2 , 4.3), (b) Yes

Problem 9. (a) (-0.0156 , 0.0956), (b) No

Problem 10. 733

Problem 11. (a) (-2.7539 , -1.2761), (b) Yes

Problem 12. (a)(0.527 , 0.625) (b) In repeated sampling 95% of the intervals produced using this method will contain the proportion of students at this university who would respond S to the question.

Problem 13. (a) (0.3602 , 0.4198) (b) In addition to the 39% who wanted major change, 19% wanted a complete overhaul of the system, leading to the total of 58% of the sample.