Chapter 1

Testing hypotheses

What is a hypothesis testing?

> "If you are different from the rest of the flock, they bite you and even reject"
>
> Vincent O'Sullivan, "The Next Room"

Hypothesis testing is another method of making inference about population parameters. In confidence intervals method we first calculate estimate based on the sample data and then make the statement about the value of the population parameter. While testing hypothesis we go from the back, namely we first make statement that we want to check and then test it with the sample data. As a result of the procedure we can support or deny the statement made.

"любит — не любит" Let's take an example from the life of the 1st year student Masha. She falls in love with her classmate Borya Beloved, but she does not know whether her love is mutual. What is she going to do? She cannot just ask him "Do you love me?". Trying her fortune on cards or with a chamomile flower does not also seem a relevant option...Well, Masha appeals for help to ...Statistics! She decides to make observations to answer her question. As she will observe the Beloved behavior, she'll get enough knowledge to test the mutuality of her love hypothesis. In order to conduct such a test she needs a threshold – an important event, which will give her enough information of interest. What an amazing fluke, the Saint Valentine's Day is coming! So, Masha builds her decision strategy: if Borya loves her, she thinks, he will inevitably present her with the Valentine's card. If he does so – Masha will happily establish herself in the hypothesis of love, otherwise – she'll reject the hypothesis and suffer from the unrequited love...

This strategy is an exact illustration of what we call a statistical test.

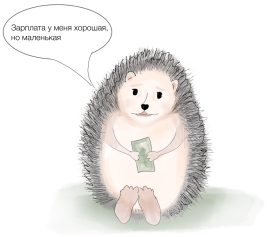
- We formulate the so-called **null hypothesis** ("Borya loves me") and the **alternative hypothesis** ("He doesn't").
- Then, we make observations (By the way, Borya did not present her with the card...).
- After that we make a thought experiment. We assume that the null hypothesis is true and assess the probability to get the observed result given the true null hypothesis. ("If he loves me, how probable is that he would not present me with the valentines' card?")
- If the estimated probability of observed data is too low ("I can't imagine he would do so if he loves me!") the null hypothesis is rejected in favor of the alternative hypothesis ("So, he does not love me!"). Otherwise – the null hypothesis is not rejected and we confirm in our belief it is true.

Note that testing hypotheses is a decision strategy based on limited amount of observed data. It does not just give you the right answer. Instead, it gives you the answer associated with some degree of confidence in it. So, mistakes are possible!

How to conduct a test?

Let's consider basic notions of testing for hypothesis about μ. Is it reasonable to believe politicians' statements about economic condition in the country? Now you've got the power to *test* their statements! No need to rely on belief of any kind.

Example 1. "Overestimated salary"



Chuvyakin is a governor of some Russian city. He states that the average monthly salary in the city is 55 000 roubles. His statement seems questionable, there is a suspicion that the real figure is lower.

We want to test this statement. Let's denote salary of a citizen by X, the true mean salary is μ. Suppose also that based on previously gathered data we know the population standard deviation to be $\sigma = 15\ 500$ roubles.

1. First, we formulate the hypotheses: H_0 (null hypothesis) and H_A (alternative hypothesis). The null hypothesis should be based on some particular number – on *a claim*.

   Here it is the statement that the mean salary is 55 000. Contrary, H_A should *express the direction of our suspicion* with regard to the statement. Chuvyakin might be overestimating the mean salary trying to show himself better in citizens' eyes. So, we suspect the true figure to be below 55 000.

   Thus,
   H_0: $\mu = 55000$
   H_A: $\mu < 55000$

We should also choose the **significance level** α. It is a threshold, a criterion to reject H_0. α is your subjective evaluation of what is a "small probability". If some event happens with probability 1% or lower you might consider it is too rare to take this possibility into consideration. Keep in mind that α also reflects your readiness to conduct the type I error. If α is not given in the problem – you should choose it yourself. The conventional values of α are: 1%, 5%, 10%.

Let's take 5% or equivalently α=0.05

2. Now we gather data. Say, we've taken a simple random sample of 40 citizens, obtained data on their salaries X, and calculated average salary to be $\bar{X} = 35000$. Recall also that σ is known to be 15,500.

3. We known that sample was obtained by SRS. $n = 40 > 30$, so, we may consider sample as large enough to state that $\bar{X} \sim N(\mu, \frac{\sigma}{\sqrt{n}})$. This will allow us to test the hypothesis.

4. Now we assume that H_0 is true and calculate z-score corresponding to the observed sample mean $\bar{X}$=35000, This will allow us to evaluate likelihood of our observations from the H_0 point of view. Recall that z-score is the number of standard deviations by which observed value lies away from its mean. Then, the higher is the absolute value of z-score, the further is the observed result from μ, assumed by H_0, so, the more unusual observations are!

   Absolute value of z-statistic reflects *how unlikely it is to get what we've actually got given H_0 is true.*

$$z_{\text{st}} = \frac{\bar{X} - \mu_0}{\frac{\sigma}{\sqrt{n}}}$$

$$z_{\text{st}} = \frac{\bar{x} - \mu_0}{\frac{\sigma}{\sqrt{n}}} = \frac{35000 - 55000}{\frac{15500}{\sqrt{40}}} = \frac{35000 - 55000}{2450.77} = -8.16$$

   So, observed value is very far (by 8 standard deviations) from what it is expected to be according to H_0.

   Note, that all calculations are done "*under null hypothesis*" – using parameters from H_0.

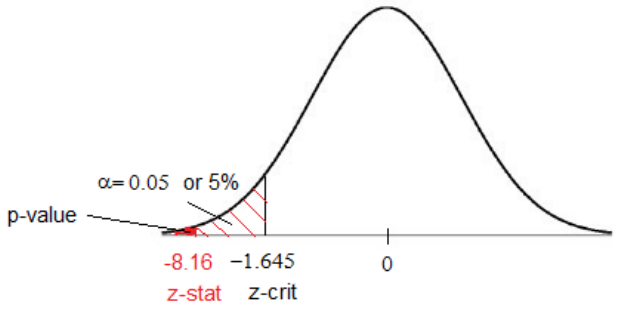
5. Now, how can we evaluate the "unusualness" of observations shown in z-statistic? Via probability! At this step we calculate the so-called p-value. **p-value** is the probability to obtain the observed or even more extreme sample (and thus the corresponding z-statistic) given H_0 *is true.* In this context "extreme" means unusual from the null hypothesis point of view, "more extreme" – means "lies even further from the hypothesized mean".

   In other words, p-value answers the question "If H_0 is true, what would be the probability to get sample mean $\bar{X}$ equal to 35000 or even less?"

   p-value $= P(\bar{X} \leqslant 35000) = P(z \leqslant -8.16) = 1,7 \cdot 10^{-16}$

6. Now we make a comparison of p-value versus α.

   p-value $= 0.00000000000000017 < \alpha = 0.05$. Thus, probability of the observed result is too low under H_0. So, if H_0 was true, it would be *too much unlikely* to get the sample mean salary of 35,000. Thus, we have enough statistical evidence to *reject the null hypothesis at 5% significance level* and not to believe Chuvyakin.



5* Note that you can reach the same conclusion without calculating the p-value. In this case you need to compare z_{st} with the so-called critical value z_{cr}. Look at the picture above. As you can see, comparison of the two areas to the left of 0 (p-value versus α) is equivalent to comparison of z values corresponding to them. After we've chosen α, we cut off the corresponding rejection area of 0,05 on the pdf curve (since our alternative hypothesis is left-sided, we take 1% from the left side of pdf curve). The value of z which cuts off this area is called z-critical. It is calculated from: $P(z < z_{\text{cr}}) = \alpha$. In our case $P(z < z_{\text{cr}}) = 0.01$ and $z_{\text{cr}} = -z_\alpha = -z_{0,01} \approx -1.645$. Since z_{st} is farther from 0 than z_{cr} (or $|z_{\text{st}}| > |z_{\text{cr}}|$) it gets into the rejection area. So, H_0 is rejected in favor of H_A at 5% significance level.

Example 2. Underestimated mean number of bureaucrats per project

Chuvyakin states that the team of employees working for him is not too big. On average, he says, a project group for each problem in the city includes no more than 25 people. An activist took a sample of 20 projects conducted last year and noticed that average declared number of participants was 37 per project. Maybe Chuvyakin is lying again...

1. We state hypothesis and choose significance level.

   Let's denote the number of participants per government project by X, μ is the true mean number of participants.
   H_0: $\mu = 25$ (always an equality even if Chuvyakin stated no more than)
   H_A: $\mu > 25$
   Let's take significance level $\alpha = 5\%$.

2. We have that $\bar{X} = 37$, $n = 20$. Suppose also historically we know the population standard deviation: $\sigma = 14$.

3. Then, we check conditions.

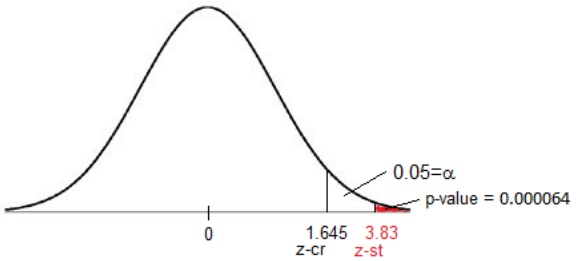
   1) Assume that the data was obtained from a simple random sample and that it is no more than 5% of population.

   2) $n = 20 < 30$, that is, not large enough for CLT to guarantee that $\bar{X}$ is normal. So, in order to conduct a test we first need to assume that $X \sim N$ (sample mean $\bar{X}$ is a sum of independent normal variables X, so, if X is normal, then, $\bar{X}$ is also normal).

4. z-statistic $= \frac{\bar{x}-\mu_0}{\frac{\sigma}{\sqrt{n}}} = \frac{37-25}{\frac{14}{\sqrt{20}}} = 3.83$

5. p-value $= P(\bar{X} > 37) = P(z > 3.83) = 0.000064$

6. p-value $= 0.000064 < \alpha = 0.05$ Thus, if H_0 was true, it would be *too much unlikely* to get a sample mean number of bureaucrats equal to 37. Thus, we have strong evidence not to believe Chuvyakin and *reject the null hypothesis* at 5% significance level.



5* Alternatively you can find $z_{\text{cr}} = z_{0,05} = 1,645$. Since $z_{\text{st}} = 3,83 > z_{\text{cr}}$, H_0 is rejected at 5% significance level.

Note, that in examples 1 and 2 we used *one-sided alternative hypotheses* ($\mu > 25$, $\mu < 55000$). This is because our suspicion about Chuvyakin's claim in both cases is directed.

Example 3. "Average yield of milk"

From one of the farms located in the region of Chuvyakin's control, there came a report stating that the average yield of milk by one cow is 12 liters per day. Chuvyakin asks his assistant to check this report. Both of them are living in the capital and don't have a clue what the average milk yield should be. So they want to test this report. Population standard deviation is known to be 5.0.

They took a sample of 25 cows and got the average of 10.8 lt.

1. Let X be the amount of milk produced by a cow in this region.
   H_0: $\mu = 12.0$
   H_A: $\mu \neq 12.0$

Note, that in this example our suspicion about the claim has no special direction. So, the *alternative hypothesis is two-sided.*

We choose $\alpha = 5\%$.

2. $n = 25$, $\bar{X} = 10.8$, $\sigma = 5$.

3. Check conditions:

   1) We assume the sample was SRS and is no more than 5% of the population.

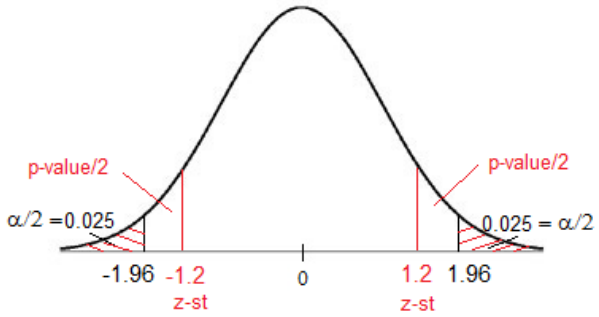
   2) Our sample is small: $n = 25 < 30$. We need *to assume that* $X \sim N$.

   3) $z_{\text{st}} = \frac{\bar{x}-\mu_0}{\frac{\sigma}{\sqrt{n}}} = \frac{10.8-12}{\frac{5.0}{\sqrt{25}}} = \frac{10.8-12}{1} = 1.2$

   4) p-value $= P(|\bar{X}| > 12.0) = P(|z| > 1.2) = 2P(z > 1.2) = 0.115 \cdot 2 = 0.23$

Note that in the case of two-sided alternative we "double" the p-value, using probability of deviation in *absolute value.* We not only count the probability to get z higher than 1.2, but add up the chance of getting the opposite result: $z < -1.2$. This is because we are interested in probability of $\bar{X}$ deviating from μ by 1.2 *in any direction.* Based on our alternative we are not interested in any particular direction of deviation.

4. p-value $= 0.23 > \alpha = 0.05$. Thus, we think that, if H_0 was true, it is quite likely to get average 10.8 and so we have no evidence against H_0. We *do not reject* H_0 at 5% significance level.



5* Alternatively, we find z_{cr}. Since we work with the 2-sided alternative, α proportion of the most unusual observations should be equivalently cut off from both sides of the distribution. So, we take $\frac{\alpha}{2}$ from both sides. Then, $z_{\text{cr}} = -z_{\frac{\alpha}{2}} = -z_{0,025} = -1,96$. Since $|z_{\text{st}}| < |z_{cr}|$, H_0 is not rejected at 5% significance level.

Here is the short strategy for conducting any test (for the full version address the end of this chapter):

1. State hypotheses and choose significance level:
   H_0: ...
   H_A : ...
   Let $\alpha = \ldots$

2. Introduce a variable and write down given statistics.

3. Check requirement and state assumptions.
4. Calculate z-statistic.
5. Find p-value.
6. Compare p-value with α (or z_{st} with z_{cr}) and conclude about the H_0.

So, in the first two examples, we've rejected the null hypothesis H_0, while in the third example, we did not reject H_0. Note that although it might sound similar, you cannot state that "the alternative hypothesis H_A is proved" (first two examples) or that "the null hypothesis H_0 is accepted" (the third example). Such wording of the answer is not applicable and will be viewed as incorrect. This is because we should always bear in mind that we base our conclusion on the limited data of a sample, and therefore it contain some level of uncertainty. That is why we only may conclude that "H_0 is rejected (in favor of alternative H_A)" or that "H_0 is not rejected/there is not enough statistical evidence to reject H_0".

Type I and Type II Errors

> I would never die for my beliefs
> because I might be wrong
>
> Bertrand Russell

As we have commented above, any conclusion based on testing procedure may be wrong. Mistakes are always possible! There are 2 types of mistakes associated with testing a hypothesis. First, we can mistakenly reject the true null hypothesis (type I error). Second, we can mistakenly approve the null hypothesis, while in fact alternative hypothesis is true (Type II error). Let's discuss both types of mistakes in detail.

Type I error

Type I error is the situation when as a result of a test, null hypothesis is rejected, although it is true.

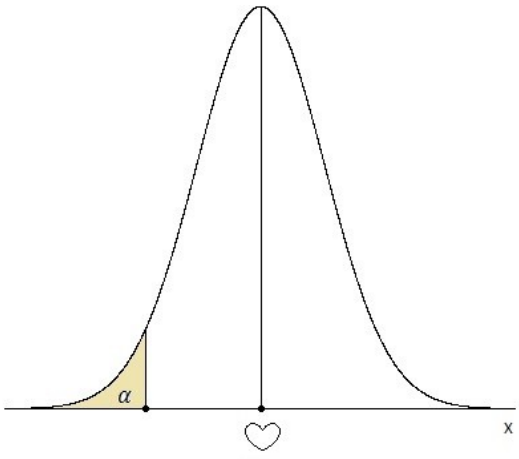
$$\text{Conduct type I error} = \text{reject the true } H_0$$

Recall the "любит — не любит" example. Masha wants to test the null hypothesis that Borya loves her. For the sake of simplicity let us assume that Borya's feelings to Masha can be observed and measured. Since Borya is subject to changes in mood, random external events and excessive self-criticism his feelings can be viewed as a random variable X. Also, let's assume X is normally distributed and μ_X is the true mean feelings.

According to the null hypothesis μ_X = love. Contrary, alternative hypothesis assumes that he is not in love, so, $\mu_X <$ love.

The picture below illustrates the distribution of X under H_0. Note, that it is centered at μ_X = love. Sending a Valentine's Day card is to the left of mean, since it

is viewed as a minimum attribute of love. The type I error will occur if Borya will not present the post card, though being in love. Consequently, Masha will mistakenly reject the love hypothesis. Though being unlikely, this event may happen. Borya can forget the current date or accidentally lose the postcard just on his way to the university, etc. The probability of such mistake is the area under the pdf curve to the left of the threshold event, that is, "the postcard presented". Probability of type I error is conventionally denoted by α.

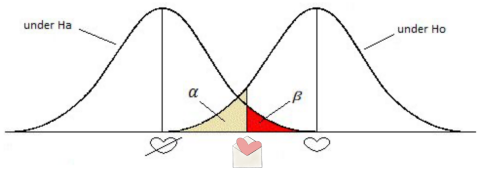


Type II error

Type II error is the situation when, null hypothesis is not rejected, although it is false.

Conduct type II error = fail to reject the false H_0

It is possible that Borya is not in love. In this case H_A is true and the feelings distribution is centered at the crossed heart. However, he can still present Masha with the card. Say, he is going to present cards to both of her best friends, so he decides to present it also to Masha just to be polite and not to discriminate against her among the other two girls. Then, Masha will mistakenly decide he is in love. The probability of such mistake is the area under the left pdf curve to the right of the postcard event (red shaded area). Probability of type II error is conventionally denoted by β.



Note that both α and β depend on the threshold event. If you change it, probabilities of both errors will change. For example, if Masha decides to choose a "higher" threshold event, to the right of current position (say, "flowers presented") α will rise,

but β will become smaller. Contrary, if we move threshold to the left, it will lower α and increase β. So, there is a tradeoff between the two types of error.

However, is it possible to reduce both α and β? The only way to do that is to increase sample size. Standard deviation of the sample statistic will then fall, and both distributions will become thinner, reducing both α and β (shaded areas to the left and to the right of the threshold).

Overall, four situations are possible in testing a hypothesis. If H_0 is true we can either reject it (Type I error, occurs with probability α) or not reject it (right decision, is taken with probability $1-\alpha$). Contrary, if H_A is true we either reject H_0 (right decision, is taken with probability $1-\beta$) or not reject H_0 (Type II error, occurs with probability β). Conventionally, the concerned probabilities are given the following names: α = significance level, $1-\alpha$ = confidence level, $1-\beta$ = power of the test.

The table below, sometimes addressed to as the confusion matrix, summarizes all the cases described.

	Not reject H_0	Reject H_0 in favor of H_A
H_0 is true	Right decision. Probability = $1-\alpha$ = confidence level.	Type I error. Probability = α = significance level
H_A is true	Type II error. Probability = β	Right decision. Probability = $1-\beta$ = power of the test

Here is the illustration of how probability of these two errors can be found.

Recall the **OVERESTIMATED SALARY** example. We test the following pair of hypotheses about the mean salary:
H_0: $\mu = 55000$
H_A: $\mu < 55000$

Type I error. Suppose someone states the following decision rule in this test: to reject the null hypothesis whenever sample mean is below 50 969 roubles.

What is the of type I error? It is the mistake that will occur if the sample mean turns out to be lower than... although H_0 is true and $\mu = 55\ 000$.

What is the probability of type I error? It is the chance to reject H_0 given that it is true:

$$P(\text{reject } H_0 | H_0 \text{ is true}) = P(\bar{X} < 50969 | \mu = 55000) =$$
$$= P\left(z < \frac{50969 - 55000}{\frac{15500}{\sqrt{40}}}\right) = P(z < -1.645) = 0.05 = \alpha$$

As you can see, probability of type one error equals the significance level α which was given in the initial example. Here we artificially used 50969 roubles as the threshold in the decision rule to explain how to find probability of type I error when α is not given, and you only have the decision rule (rejection rule).

Type II error. Suppose that the true value of mean salary in the region is 40000 and you use the same decision rule to test the stated above hypothesis.

What is the type II error? It is the situation when we will mistakenly fail to reject H_0: $\mu = 55000$ when in fact H_A is true and $\mu < 55000$.

What is the probability of type II error β?

$$\beta = P(\text{not to reject } H_0 | H_A \text{ is true}) = P(\bar{X} > 50969 | \mu = 40000) =$$
$$= P\left(z > \frac{50969 - 40000}{\frac{15500}{\sqrt{40}}}\right) = P(z > 4.476) = 3.8 \cdot 10^{-6}$$

What is the power of the test? Power of the test $= 1 - \beta = 1 - 0.0000038 = 0.9999962$

In this chapter we are going to discuss four types of statistical tests: about the population mean μ, the population proportion p, difference in population means $\mu_1 - \mu_2$ and difference in population proportions $\pi_1 - \pi_2$.

Let's start the journey!

Testing hypotheses for population parameters (μ and p)

Tests for the population mean μ

Population standard deviation σ is known

$$z_{\text{st}} = \frac{\bar{X} - \mu_0}{\frac{\sigma}{\sqrt{n}}}$$

Note that this is just one realization of the more general formula for any test statistic:

$$\text{Test-statistic} = \frac{\text{Estimate} - \text{Parameter}_0}{\text{SE}_{\text{Estimator}}}$$

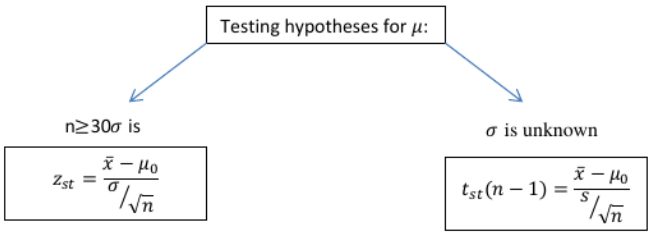
Population standard deviation σ is unknown

Of course in real life it is almost always that σ is unknown. When we replace σ by s in the formula $z = \frac{\bar{x}-\mu}{\frac{\sigma}{\sqrt{n}}}$ the distribution is no more normal. As you know from the chapter on confidence intervals, $\frac{\bar{x}-\mu}{\frac{s}{\sqrt{n}}} \sim t(n-1)$.

Testing hypothesis then, is analogical to the case with known σ but the statistic used is:

$$t_{\text{st}}(n-1) = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}}$$

The algorithm of choosing a statistic for testing hypothesis about population mean μ is below:



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 check whether the normality assumption is reasonable based on sample distribution. If this is impossible you should state the corresponding assumption.

Conditions: Sample is SRS $n > 30$, otherwise $X \sim N$

Testing hypotheses for the population proportion

The strategy for testing hypothesis for population proportion p is essentially the same as for population mean.

First, we state the null hypothesis, which fixes some particular value of proportion p_0 which is aimed to be tested: H_0: $p = p_0$.

We also state the alternative hypothesis. Depending on the direction of our suspicion it may be one of the three types: H_A: $p > p_0$, H_A: $p < p_0$ (one-sided alternatives) and H_A: $p \neq p_0$ (two-sided alternative).

We choose the value of significance level α.

Then, we take a sample of size n and calculate sample proportion $\hat{p}$.

Then, we need to evaluate how far from the hypothesized value of p_0 has $\hat{p}$ fallen. So, we need a statistic!

As we've learned from the previous chapter, given large enough number of observations, sample proportion is normally distributed with the following parameters: $\hat{p} \sim N(p, \sqrt{\frac{p(1-p)}{n}})$. Sample size should be enough to assure that $\hat{p}$ is approximately normal by CLT. So, you should check that

Assuming the null hypothesis is true, $\hat{p} \sim N(p_0, \sqrt{\frac{p_0(1-p_0)}{n}})$.

Then, $\frac{\hat{p}-p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} = z \sim N(0,1)$. So, the statistic used is:

$$z_{\text{st}} = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$$

Value of z-statistic tells us how far from the assumed mean value p_0 has the sample proportion fallen. That is, by how may standard deviations has $\hat{p}$ deviated from p_0. If the value is too big – we reject the null hypothesis in favor of the alternative.

The decision rule is the same as for tests on population means: you either compare z-statistic and z-critical (if z_{st} exceed z_{cr} in absolute value – reject H_0), or compare p-value with α (reject H_0 when p-value $< \alpha$).

Note that *this is the only formula* used for testing statistic for p. It is only used when sample is large enough (tested by $n\hat{p}$ and $n(1-\hat{p})$ values) and uses only normal distribution (no t-distribution for proportions!).

Check conditions:

Sample is SRS $n\hat{p} \geqslant 5$, $n(1-\hat{p}) \geqslant 5$

Testing hypotheses about difference of parameters

Testing hypotheses for the difference in population means $\mu_1 - \mu_2$

In many cases we are interested in the difference between two populations. For example, we might be interested in whether this year students have passed AP exam better than the last year students *on average*. To test that, we need to assess how far does the observed difference lie from zero.

Thus, to answer the question "is there difference in means?" we need a zero difference in null hypothesis: H_0: $\mu_1 - \mu_2 = 0$ or equivalently H_0: $\mu_1 = \mu_2$. All the formulas below assume the above H_0.

What statistic should be used for testing hypothesis on difference between μ_1 and μ_2?

Known population standard deviations σ_1, σ_2

Well, we already know that for large enough samples: $\bar{x_1} - \bar{x_2} \sim N(\mu_1 - \mu_2, \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}})$

Under H_0: $\bar{x_1} - \bar{x_2} \sim N(0, \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}})$.

Hence, the statistic $z_{\text{st}} = \frac{\bar{x_1} - \bar{x_2} - 0}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$ will give us the measure of plausibility of null hypothesis (closer to zero means less deviation, then, more plausible H_0)

$$z_{\text{st}} = \frac{\bar{x_1} - \bar{x_2} - 0}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

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

Given that σ_1 and σ_2 are unknown (which is usually the case in real life problems) we are forced to replace it with s_1 and s_2. We've already shown in previous chapter that given large enough samples n_1 and n_2: $\frac{\bar{x_1} - \bar{x_2} - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \sim t\,(k-1)$, $k = \min\{n_1, n_2\}$.

Thus, the statistics used in test is:

$$t_{\text{st}}\,(k-1) = \frac{\bar{x_1} - \bar{x_2} - 0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}},\ k = \min(n_1, n_2)$$

Note that your graphic calculator uses another formula for degrees of freedom. If you use the results from the calculator in solving a problem – you should indicate df used there.

These formulas are applicable when both samples are large enough (so that we can apply CLT). For small samples we need to make sure that X_1 and X_2 are normal. Be sure to check/assume that. You should also make sure that the samples of X_1 and X_2 are independent.

Conditions:

Both samples are SRS
$n_1 \geqslant 30$ or $X_1 \sim N$, $n_2 \geqslant 30$ or $X_2 \sim N$
X_1 and X_2 are independent

Unknown population standard deviations, equal variances assumption: $\sigma_1 = \sigma_2$

When the assumption of equal variances $\sigma_1 = \sigma_2 = \sigma$ is held distribution of difference in sample means has the following parameters:

$\bar{x_1} - \bar{x_2} \sim N(\mu_1 - \mu_2, \sigma \cdot \sqrt{\frac{1}{n_1} + \frac{1}{n_2}})$. Then, $\frac{\bar{x_1} - \bar{x_2} - (\mu_1 - \mu_2)}{\sigma \cdot \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} = z \sim N(0, 1)$

We use s-pooled to estimate the unknown population standard deviation (reread the 3.2 section of the chapter on confidence intervals for explanation): $\hat{\sigma} = s_p = \sqrt{\frac{s_1^2 \cdot (n_1 - 1) + s_2^2 \cdot (n_2 - 1)}{n_1 + n_2 - 2}}$

Replacing σ by s_p requires to use t-distribution instead of normal: $\frac{\bar{x_1} - \bar{x_2} - (\mu_1 - \mu_2)}{s_p \cdot \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t(n_1 + n_2 - 2)$.

Under the assumption that the null hypothesis is true we have: $\mu_1 - \mu_2 = 0$. Thus, the following statistic is used:

$$t_{\text{st}}\,(n_1 + n_2 - 2) = \frac{\bar{x_1} - \bar{x_2} - 0}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

Here the set of assumptions is the same as in the previous case *plus* assumptions that $\sigma_1 = \sigma_2$

Conditions:

Both samples are SRS
$n_1 \geqslant 30$ or $X_1 \sim N$, $n_2 \geqslant 30$ or $X_2 \sim N$
X_1 and X_2 are independent
$\sigma_1 = \sigma_2$

Matched/paired samples case

Sometimes a problem requires to compare the means of two matched samples (for explanation of what is a matched sample read paragraph 3.3 of the previous chapter). The null hypothesis assumes no mean difference: H_0: $\mu_{\text{diff}} = 0$.

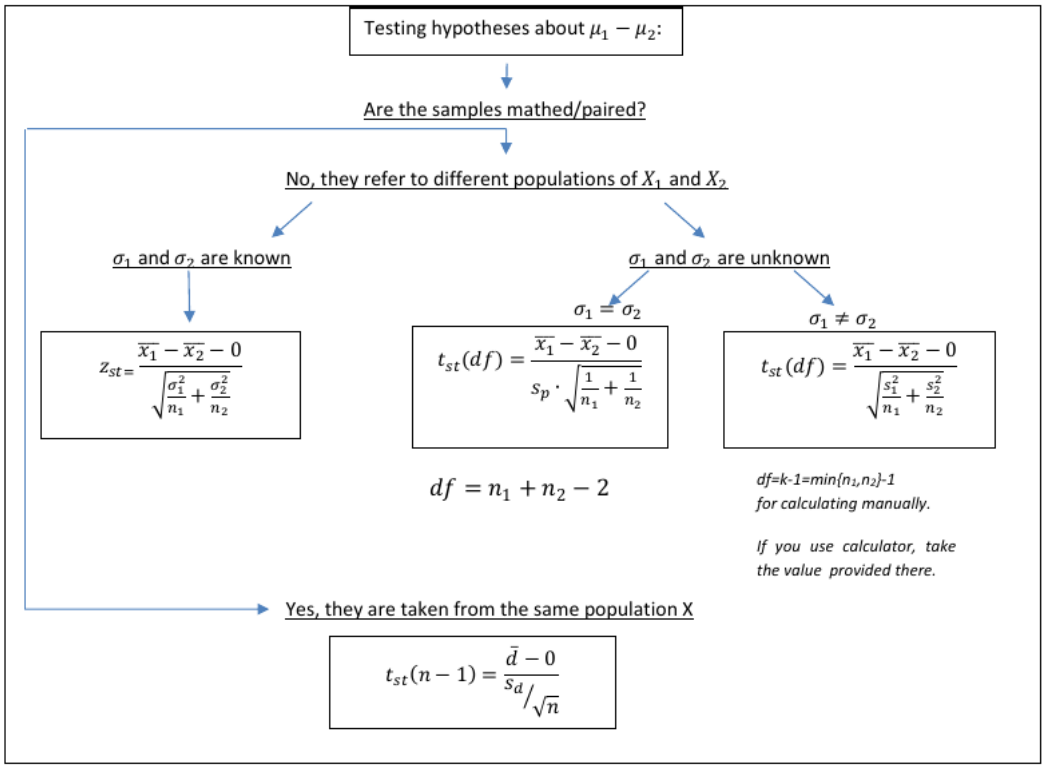
The two samples are joined to produce one – the sample of individual differences: $d_i = X_{1i} - X_{2i}$. The new sample mean and the sample standard deviation are denoted by $\bar{d}$ and s_d. Then, we come to the same problem as in 1.2 – testing the hypothesis about population mean given the true standard deviation is unknown. So, we use the following statistic:

$$t_{\text{st}} = \frac{\bar{d} - 0}{\frac{s_d}{\sqrt{n}}}$$

Here you need to make sure that the matched sample is SRS and that it is large enough.

Conditions:

Sample is SRS $n \geqslant 30$ or $\bar{d} \sim N$



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!

$$\text{Test-statistic} = \frac{\text{Estimate} - \text{Parameter}_0}{\text{SE}_{\text{Estimator}}}$$

Testing hypotheses for the difference in population proportions $(p_1 - p_2)$

In the previous chapter it was shown that difference in population proportions has the following normal distribution given that both samples are large enough: $\hat{p}_1 - \hat{p}_2 \sim N\left(p_1 - p_2, \frac{p_1 \cdot (1-p_1)}{n_1} + \frac{p \cdot (1-p_2)}{n_2}\right)$

According to H_0: p_1-$p_2 = 0$. Since proportions are equal, let's denote them by $p = p_1 = p_2$. Assuming H_0 is true we get: $\hat{p}_1 - \hat{p}_2 \sim N\left(0, p\,(1-p) \cdot \left(\frac{1}{n_1} + \frac{1}{n_2}\right)\right)$.

Deducting the mean and dividing by standard deviation we get the standard normal variable: $z = \frac{\hat{p}_1 - \hat{p}_2 - 0}{\sqrt{p(1-p) \cdot \frac{1}{n_1} + \frac{1}{n_2}}}$

This looks like a statistic for testing hypothesis! However, how do we get p to put into the denominator? Of course we cannot calculate it, but it can be estimated based on the sample data. Under the null hypothesis population means of both p_1 and p_2 are the same and equal to p. Therefore it is reasnoble to estimate it on the joint sample of n_1 and n_2 to produce the best estimate. Given that $\hat{p}_1 = \frac{m_1}{n_1}$ and $\hat{p}_2 = \frac{m_2}{n_2}$ we get:

$$\hat{p} = \frac{m_1 + m_2}{n_1 + n_2} = \frac{\hat{p}_1 . n_1 + \hat{p}_2 . n_2}{n_1 + n_2}$$

Thus, the statistic needed to test the hypothesis about equal proportions is:

$$z_{\text{st}} = \frac{\hat{p}_1 - \hat{p}_2 - 0}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

Note that *this is the only formula* of testing statistic for difference in proportions. It is only used when sample is large enough and uses only normal distribution (no t-distribution for proportions!). You should also make sure that the two samples are independent.

Conditions:

Both samples are SRS $n_1\hat{p}_1 > 5$ and $n_1(1-\hat{p}_1) > 5$, $n_2\hat{p}_2 > 5$ and $n_2(1-\hat{p}_2) > 5$ Samples are independent

Below we present the example of how to test a hypothesis on difference in proportions.

The full strategy for testing a hypothesis

Here is the strategy for testing any hypotheses from the list presented above.

Step 0. What is the problem about?
About the population mean μ, population proportion p, difference in population means $\mu_1 - \mu_2$ or difference in population proportions $p_1 - p_2$? Based on that, introduce variables you are going to use: "Let X be..." or "Let p be..." or "Let X_1 and X_2 be..." or "Let p_1 and p_2 be..."

Step 1. State the hypotheses and choose α
The null hypothesis H_0 always contains equality (=). Otherwise we would be unable to calculate any specific statistic for the observations (e.g. what to put as μ_0 into the formula of z_{st}?).

Alternative hypothesis H_A should contain one of the following signs: $<$, $>$, $\neq$.

You choose the alternative according to the direction of your suspicion about the statement in H_0. Depending on the option chosen there are different ways to calculate p-value and to compare z_{st} versus z_{cr} or t_{st} versus t_{cr} (for full explanation see the three examples in the beginning of the Chapter).

If significance level α is not given in the problem – choose it yourself: 1%, 5% or 10%.

Step 2. Calculate (and write down) the sample statistics.
E.g. $\bar{X} = \ldots$, $s = \ldots$, $n = \ldots$.

Step 3. Choose the formula and state the assumptions/check conditions.

1. SRS. Assume the sample(-s) is a SRS and is no more than 5% of the population. This assumption is always needed!

2. Normality. Check sample size.

   - if n is large ($n \geqslant 30$ for mean and $np \geqslant 5$, $n(1-p) \geqslant 5$ for proportion), normality is assured by CLT
   - n is small – check/assume that X (or X_1 and X_2) is normal. For proportions calculated on small samples no z-test is possible!

3. (for mean only) Sigma. Is σ known? Based on that you can choose the right formulas to apply. If yes – apply z-statistic, if no – use t-statistic.

4. (for difference in means or in proportions) Independence. Assume the two samples are independent.

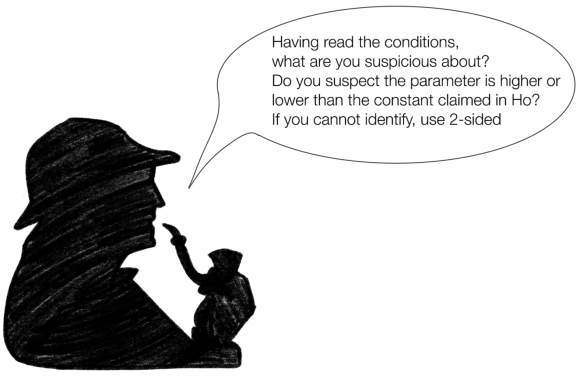
5. (for the formula with s_p). Equal standard deviations. Is it reasonable to assume that $\sigma_1 = \sigma_2$? If so, *write down this assumption.*

Step 4. Calculate z- or t-statistic
(find degrees of freedom: df=...)

Step 5. Based on the type of hypothesis and α chosen find p-value or z-critical/t-critical

Step 6. Compare p-value with α or z_{st} versus z_{cr} or t_{st} versus t_{cr}

Step 7. State your conclusion at given significance level (e.g. "H_0 is rejected at 5% significance level") and interpret the result (e.g. "Thus, there is no evidence that mean consumption exceeds 200$").



Example

Suppose that early in Chuvyakin's election campaign a telephone poll of 800 registered voters shown 460 in favor of him. Just before the election day a second poll shown only 520 of 1000 registered voters in favor of him. At 10% significance level is there sufficient evidence that Chuvyakin's popularity has decreased?

Solution:

Step 0. We are asked about the difference in proportions. Let p_1 and p_2 be the true proportions of voters in favour of Chubyakin early in election and campaign and just before the election.

Step 1. H_0: $p_1 - p_2 = 0$ proportions in favour of Chuvyakin are the same
H_A: $p_1 - p_2 > 0$ Chuvyakin's popularity has decreased
Let's take $\alpha = 0.1$ or 10%

Step 2. $\hat{p}_1 = \frac{460}{800} = 0.575$ $\hat{p}_2 = \frac{520}{1000} = 0.520$

Step 3. *(there is only 1 formula for difference in proportions)*

- We assume that both samples were SRS (simple random samples) of the electorate.
- $n_1 p_1 \geqslant 5$, $n_1(1-p_1) \geqslant 5$, $n_2 p_2 \geqslant 5$, $n_2(1-p_2) \geqslant 5$. Thus, we can use that both p_1 and p_2 are approximately normal.
- We assume that the samples are independent from each other.

Step 4. A tricky moment! Under H_0 p is the same in both populations! We should find an estimate for it.

$\hat{p} = \frac{\hat{p}_1 . n_1 + \hat{p}_2 . n_2}{n_1 + n_2} = \frac{460+520}{800+1000} = \mathbf{0.544}$

$z_{\text{st}} = \frac{p_1 - p_2 - 0}{\sqrt{p(1-p)}\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} = \frac{(0.575-0.520) - 0}{\mathbf{0.0236}} = \frac{0.055-0}{0.0236} = \mathbf{2.33}$

Step 5. $p - value = P(z > 2.33) = 0.0099$

Step 6. Since $0.0099 < 0.1$, we conclude that at 10% significance level we reject H_0 of equality of proportions.

Step 7. We have evidence to think that the popularity of Chuvyakin has fallen.

You must be able to reproduce even being half-awake

- The full strategy of testing a hypothesis
  - Step 0. What is the problem about? "let it be" song.
  - Step 1. State the hypotheses and choose α
  - Step 2. Calculate (and write down) the sample statistics. E.g. $\bar{X} = \ldots$, $s = \ldots$, $n = \ldots$
  - Step 3. Choose formula and state assumptions.
  - Step 4. Calculate z- or t-statistic (find degrees of freedom: df $= \ldots$)
  - Step 5. Based on the type of hypothesis and α chosen find p-value or z-critical/t-critical.
  - Step 6. Compare p-value with α or z_{st} versus z_{cr} or t_{st} versus t_{cr}.
  - Step 7. State your conclusion at given significance level.
- Conduct type II error = fail to reject the false H_0
- Conduct type I error = reject the true H_0

Calculator Box

$z_{\text{st}} = \frac{\bar{x}-\mu_0}{\frac{\sigma}{\sqrt{n}}}$: `TEST` → `z` → `1-s` (stands for 1 sample) → specify the type of H_A ($>$, $<$, $\neq$), enter values of statistics → `Exe`

$t_{\text{st}}(n-1) = \frac{\bar{x}-\mu_0}{\frac{s}{\sqrt{n}}}$: `TEST` → `t` → `1-s` → indicate H_A type, 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.

$z_{st} = \frac{\bar{x_1}-\bar{x_2}-0}{\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}}$: `TEST` → `z` → `2-s` (stands for 2 samples)→ indicate H_A type, values of statistics → `Exe`

$t_{\text{st}}(df) = \frac{\bar{x_1}-\bar{x_2}-0}{\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}}$: `TEST` → `t` → indicate H_A type, values of statistics, choose "Off" for the "Pooled" option → `Exe`

Note that calculator return different number of degrees of freedom than $\min\{n_1, n_2\} - 1$. Using results from calculator, specify df indicated there!

t_{st} (df) $= \frac{\bar{x_1}-\bar{x_2}-0}{s_p \cdot \sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}$: `TEST` → `t` → indicate H_A type, values of statistics, choose "Off" for the "Pooled" option → `Exe`

$t_{\text{st}}\,(n-1) = \frac{\bar{d}-0}{\frac{s_d}{\sqrt{n}}}$ is calculated in the same way as t-statistic for 1 sample. You should just put values of d_i into a List, then: `TEST` → `t` → `1-s` → Data: `List`, indicate H_A type → `Exe`

$z_{\text{st}} = \frac{\hat{p}-p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$: `TEST` → `z` → `1-p` → indicate H_A type, values of statistics (and n where $\hat{p} = \frac{x}{n}$) → Exe

$z_{\text{st}} = \frac{\hat{p}_1-\hat{p}_2-0}{\sqrt{\hat{p}(1-\hat{p})}\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}$:TEST → z → 2-p → indicate H_A type, values of statistics → Exe

Top secret information

Note that testing hypothesis with one-sided alternative is equivalent to conducting one-sided confidence interval to test a statement. Analogously, two-sided alternative will result in the same conclusion as two-sided confidence interval.

Let's take example 2. In order to test the statement that mean number of people per project does not exceed 25, we will construct a lower bound for μ to check whether it is below 25:

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

$\mu > 37 - 1.645\frac{14}{\sqrt{20}} \approx 31.85$.

Thus, we are 95% confident, that μ belongs to $[31.85, +\infty)$. In other words, we are confident that mean is above 32 persons, which exceeds 25. Since $25 \notin [31.85, +\infty)$, we reject the statement $\mu = 25$.

Now let's address example 3. The two-sided interval for μ would be: $\mu = \bar{X} \pm z_{\text{critical}} \cdot \frac{\sigma}{\sqrt{n}} = 10.8 \pm 1.96 \cdot \frac{5}{\sqrt{25}} = 10.8 \pm 1.96$.

We are 95% confident that $\mu \in [8.84, 12.76]$. Since $12 \in [8.84, 12.76]$, statement $\mu = 12$ is not rejected.

Thus, using $(1-\alpha)$ confidence intervals we come to the same conclusions as testing hypotheses with corresponding alternative at α significance level.

Sample AP problems with solutions

Problem 1. AP 2015 №4

A researcher conducted a medical study to investigate whether taking a low-dose aspirin reduces the chance of developing colon cancer. As part of the study, 1,000 adult volunteers were randomly assigned to one of two groups. Half of the volunteers were assigned to the experimental group that took a low-dose aspirin each day, and the other half were assigned to the control group that took a placebo each day. At the end of six years, 15 of the people who took the low-dose aspirin had developed colon cancer and 26 of the people who took the placebo had developed colon cancer. At the significance level $\alpha = 0.05$, do the data provide convincing statistical evidence that taking a low-dose aspirin each day would reduce the chance of developing colon cancer among all people similar to the volunteers?

Solution
Step 0. Let p_1 and p_2 be the population proportions of adults similar to those in the study who would have developed a colon cancer among those who take and those who do not take a low-dose aspirin, correspondingly.
Step 1. $H_0 : p_1 - p_2 = 0$, $H_a : p_1 - p_2 < 0$

$$\alpha = 0.05$$

Step 2. We have: $\hat{p}_1 = \frac{15}{500} = 0.03$, $\hat{p}_2 = \frac{26}{500} = 0.052$, $n_1 = n_2 = 500$.
Step 3.

1) The sample of 1,000 was made up of volunteers. However, we defined the population as all people similar to those in the study. So, we can view the dataset as a random sample from adults similar to the volunteers taken.The sample of 1,000 was made up of volunteers. However, we defined the population as all people similar to those in the study. So, we can view the dataset as a random sample from adults similar to the volunteers taken.

2) $n_1\hat{p}_1 = 15 > 5$, $n_1(1 - \hat{p}_1) = 485 > 5$, $n_2\hat{p}_2 = 26 > 5$, $n_2(1 - \hat{p}_2) = 474 > 5$. Then, the sample sizes are large enough to ensure normality of $\hat{p}_1$ and $\hat{p}_2$ by CLT.

3) Since volunteers were randomly assigned to the two groups, we can say that $\hat{p}_1$ and $\hat{p}_2$ are independent.

Step 4. $z_{\text{st}} = \frac{\hat{p}_1 - \hat{p}_2 - 0}{\sqrt{\hat{p}(1-\hat{p})}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$, $p = \frac{15+26}{500+500} = 0.041$.

$$z_{\text{st}} = \frac{(0.03 - 0.052) - 0}{\sqrt{0.041(1-0.041)}\sqrt{\frac{1}{500} + \frac{1}{500}}} \approx -1.754$$

Step 5. p-value $= P(z < -1.754) \approx 0.040$
Step 6. Since p-value $< \alpha$, H_0 is rejected at 5% significance level[a].
Step 7. Thus, the data provides convincing statistical evidence that taking a low-dose aspirin each day reduces the chance of developing colon cancer among

all people similar to the volunteers.

[a]Note, that if you choose $\alpha = 0.01$, p-value$> \alpha$, and H_0 would not be rejected at 1% significance level.

Problem 2. AP 2014 №5 A researcher conducted a study to investigate whether local car dealers tend to charge women more than men for the same car model. Using information from the county tax collector's records, the researcher randomly selected one man and one woman from among everyone who had purchased the same model of an identically equipped car from the same dealer. The process was repeated for a total of 8 randomly selected car models. The purchase prices and the differences (woman – man) are shown in the table below. Summary statistics are also shown.

Dotplots of the data and the differences are shown below.

Do the data provide convincing evidence that, on average, women pay more than men in the county for the same car model?

Solution

Step 0. This is a matched sample, since the datasets of prices for men and women are gathered for the same car models: each pair corresponds to one of the eight models.

Let X_{diff} represent the difference in prices, the amount by which women pay more than men, purchased the same car model.

Step 1. $H_0 : \mu_{\text{diff}} = 0$

$H_a : \mu_{\text{diff}} > 0$

Let's choose $\alpha = 0.01$

Step 2. Based on the table data:$\bar{X}_{\text{diff}} = 585$, s_{diff}=530.71, n=8.

Step 3. 1). We are stated that the data on prices was taken randomly from among everyone who had purchased the same model. The assumption of random sample of X_{diff} is satisfied.

2). The sample size n<30 is not large enough to ensure normality of $\bar{X}_{\text{diff}}$ by CLT. However, the dotplot for X_{diff} reveals no significant deviations from normal distribution, it is symmetric and has no outliers. So normality of X_{diff} can be assumed.

Step 4. $t_{\text{st}}\,(\text{df}) = \frac{\bar{X}_{\text{diff}} - \mu_{\text{diff}}}{\frac{s_{\text{diff}}}{\sqrt{n}}} = \frac{585-0}{\frac{530.71}{\sqrt{8}}} \approx 3.118$, $\text{df} = n - 1 = 7$.

Step 5. p-value= $P(t\,(7) > 3.118) \approx 0.0085$

Step 6. Since p-value $< \alpha$, H_0 is rejected at 1% significance level.

Step 7. Thus, we have enough evidence that, on average, women pay more than men in the county for the same car model.

Problem 3. AP 2013 №5

Psychologists interested in the relationship between meditation and health conducted a study with a random sample of 28 men who live in a large retirement community. Of the men in the sample, 11 reported that they participate in daily meditation and 17 reported that they do not participate in daily mediation.

The researchers wanted to perform a hypothesis test of

$$H_0 : p_m - p_c = 0$$

$$H_a : p_m - p_c < 0$$

Where p_m is the proportion of men with high blood pressure among all the men in the retirement community who participate in daily meditation and p_c is the proportion of men with high blood pressure among all the men in the retirement community who do not participate in daily meditation.

(a) If the study where to provide significant evidence against H_0 in favor of H_a, would it be reasonable for the psychologists to conclude that daily meditation causes a reduction in blood pressure for men in the retirement community? Explain why or why not.

Solution

(a) It would not be reasonable to conclude that daily meditation causes a reduction in blood pressure for men in the retirement community. The test is based on data from the observational study, which does not allow cause-and-effect conclusions, but can only reveal a relationship. It is possible that those in the community who practice meditation differ from those who do not practice it in some important characteristics which are related with blood pressure (confounding factors). If the study was an experiment, we would be able to make decisions about the effect of meditation on blood pressure.

(b) The use of normal approximation for $\hat{p}_m - \hat{p}_c$ is based on the three assumptions: the samples should be random, large enough and independent. In this case the "large enough" assumption is not satisfied: $n_m\hat{p}_m = 0 < 5$.

(c) The observed value of $\hat{p}_m - \hat{p}_c$ is $\frac{0}{11} - \frac{8}{17} \approx -0.471$. Given H_0 is true, such an extreme (or even more extreme) observation might occur in only $\frac{76}{10\ 000} = 0.0076$ proportion of cases. This is the simulated p-value. Since it is less than 0.01, the null hypotheses of equality of proportions would be rejected in favor of alternative even at 1% significance level. Thus, we conclude that proportion of people with high blood pressure is lower among those in retirement community who practice daily meditation. High blood pressure and meditation are adversely related for people in this community.

Problem 4. AP 2012 №5

A recent report stated that less than 35 percent of the adult residents in a certain city will be able to pass a physical fitness test. Consequently, the city's Recreation Department is trying to convince the City Council to fund more physical fitness programs, but the council is facing budget constraints and is skeptical of the report. He will fund the programs only if the Recreation Department can provide convincing evidence that the report is true.

The Recreation Department plans to collect data from a sample of 185 adult residents in the city. A test of significance will be conducted at a significance level of

$\alpha = 0.05$ for the following hypotheses.
H_0: $p = 0.35$
H_a: $p < 0.35$
where p is the proportion of adult residents in the city who are able to pass the physical fitness test.

(a) Describe what a Type II error would be in the context of the study, and also describe a consequence of making this type of error.

(b) The Recreation Department recruits 185 adult residents who volunteer to take the physical fitness test. The test is passed by 77 of the 185 volunteers, resulting in a p-value of 0.97 for the hypotheses stated above. If it was reasonable to conduct a test of significance for the hypotheses stated above using the data collected from the 185 volunteers, what would the p-value of 0.97 lead you to conclude?

(c) Describe the primary flaw in the study described in part (b), and explain why it is a concern.

Solution:

(a) Type II error occurs when we fail to reject the false alternative hypothesis. In the context of the study, a Type II error means **failing to reject** the null hypothesis that 35 percent of adult residents in the city are able to pass the test **when**, in reality, **less than 35 percent** are able.

The consequence of this error is that the council would **not fund** the program, and the city would continue to have a smaller proportion of physically fit residents.

(b) Because the p-value of 0.97 is so large and larger than $a = 0.05$, we **fail to reject** the null hypothesis. There is **no** convincing **evidence** that the **proportion** of adult residents in the city who are able to pass the physical fitness test **is less than 0.35**. (context of the problem)

After all, the sample proportion of $p = 0.416$ is actually higher than 0.35, which is in the opposite direction of the alternative hypothesis.

(c) This is not a **randomly selected sample** because the sample was selected by recruiting **volunteers**. It seems reasonable to think that volunteers would be **more physically fit** than the population of city adults as a whole. Therefore, the *sample* proportion will likely **overestimate** the *population* proportion of adult residents in the city who are able to pass the physical fitness test.

Problem 5. AP 2011 №4

High cholesterol levels in people can be reduced by exercise, diet, and medication. Twenty middle-aged males with cholesterol readings between 220 and 240 milligrams

per deciliter (mg/dL) of blood were randomly selected from the population of such male patients at a large local hospital. Ten of the 20 males were randomly assigned to group A, advised on appropriate exercise and diet, and also received a placebo. The other 10 males were assigned to group B, received the same advice on appropriate exercise and diet, but received a drug intended to reduce cholesterol instead of a placebo. After three months, posttreatment cholesterol readings were taken for all 20 males and compared to pretreatment cholesterol readings. The tables below give the reduction in cholesterol level (pretreatment reading minus posttreatment reading) for each male in the study.

Group A (placebo)

Group B (cholesterol drug)

Do the data provide convincing evidence, at the $\alpha = 0.01$ level, that the cholesterol drug is effective in producing a reduction in mean cholesterol level beyond that produced by exercise and diet?

Solution

Step 0. Let X_A and X_B represent Cholesterol levels of people is groups A and B, correspondingly.

Step 1. $H_0: \ \mu_A - \mu_B = 0$ (drug is not effective)

$H_a: \ \mu_A - \mu_B < 0$ (drug is effective)

$\alpha = 0.01$

Step 2. Based on the table data: $\bar{X}_A = 10.2, \ \bar{X}_B = 16.4, \ s_A$=7.66, s_B=9.4, $n_A = n_B = 10$.

Step 3.

1. The 20 observations are said to be randomly selected from the population.
2. The sample sizes (< 30) are not large enough to ensure normality of $\bar{X}_A$ and$\bar{X}_B$ by CLT. The dotplots for X_A and X_B are slightly skewed, distribution of X_B possibly has an outlier.

   However, there is no strong evidence against the hypothesis that both X_A and $X_B$are taken from normal distribution.
3. Since people were randomly assigned to the groups, we can say that X_A and X_B are independent.

$t_{\text{st}}\,(\text{df}) = \frac{\bar{X_A} - \bar{X_B} - 0}{\sqrt{\frac{s_A^2}{n_A} + \frac{s_B^2}{n_B}}} = \frac{(10.2 - 16.4) - 0}{\sqrt{\frac{7.66^2}{10} + \frac{9.4^2}{10}}} \approx -1.617$, $\text{df} \approx 17.295$ (taken from calculator).

Step 5. p-value $= P(t\,(17.3) < -1.617) \approx 0.062$

Step 6. Since p-value $> \alpha$, H_0 is not rejected at 1% significance level.

Step 7. Thus, we have no enough evidence that the cholesterol drug is effective in producing a reduction in mean cholesterol level beyond that produced by exercise and diet.

Problem 6. AP 2009 №5

For many years, the medically accepted practice of giving aid to a person experi-

encing a heart attack was to have the person who placed the emergency call administer chest compression (CC) plus standard mouth-to-mouth resuscitation (MMR) to the heart attack patient until the emergency response team arrived. However, some researchers believed that CC alone would be a more effective approach.

In the 1990s a study was conducted in Seattle in which 518 cases were randomly assigned to treatments: 278 to CC plus standard MMR and 240 to CC alone. A total of 64 patients survived the heart attack: 29 in the group receiving CC plus standard MMR, and 35 in the group receiving CC alone. A test of significance was conducted on the following hypotheses.

H_0: The survival rates for the two treatments are equal.

H_A: The treatment that uses CC alone produces a higher survival rate.

This test resulted in a p-value of 0.0761.

(a) Interpret what this p-value measures in the context of this study.

(b) Based on this p-value and study design, what conclusion should be drawn in the context of this study? Use a significance level of $\alpha = 0.05$.

(c) Based on your conclusion in part (b), which type of error, Type I or Type II, could have been made? What is one potential consequence of this error?

Solution

(a) The p-value informs us that if effectiveness of the two treatments was the same, and experiment was repeated many times, approximately 7.61% of them would result in a difference in proportions equal or higher than what is observed.

(b) Since p-value $> \alpha$, H_0 is not rejected at 5% significance level. There is no sufficient evidence that CC alone is more effective than CC plus standard MMR.

(c) Since we did not reject H_0, Type II error might have occurred. It is possible that in fact H_A is true. In that case we had mistakenly decided that CC alone might be equally effective as CC plus standard MMR, and continue applying less effective combined therapy, while CC alone would save more lives.

Problem 7. AP 2009 Form B №5

A bottle-filling machine is set to dispense 12.1 fluid ounces into juice bottles. To ensure that the machine is filling accurately, every hour a worker randomly selects four bottles filled by the machine during the past hour and measures the contents. If there is convincing evidence that the mean amount of juice dispensed is different from 12.1 ounces or if there is convincing evidence that the standard deviation is greater than 0.05 ounce, the machine is shut down for recalibration. It can be assumed that the amount of juice that is dispensed into bottles is normally distributed.

During one hour, the mean number of fluid ounces of four randomly selected bottles was 12.05 and the standard deviation was 0.085 ounce.

(a) Perform a test of significance to determine whether the mean amount of juice dispensed is different from 12.1 fluid ounces. Assume the conditions for inference are met.

(b) To determine whether this sample of four bottles provides convincing evidence that the standard deviation of the amount of juice dispensed is greater than 0.05 ounce, a simulation study was performed. In the simulation study, 300 samples, each of size 4, were randomly generated from a normal population with a mean of 12.1 and a standard deviation of 0.05. The sample standard deviation was computed for each of the 300 samples. The dotplot below displays the values of the sample standard deviations.

Use the results of this simulation study to explain why you think the sample provides or does not provide evidence that the standard deviation of the juice dispensed exceeds 0.05 fluid ounce.

Solution

1. Step 0. Let X be the amount of juice dispensed by the bottle-filling machine.

Step 1. $H_0 : \mu = 12.1$
$H_a : \mu \neq 12.1$
Let's choose $\alpha = 0.05$
Step 2. $X \sim N$, $\bar{X} = 12.05$, $s = 0.085$, $n = 4$.
Step 3. The conditions for inference are assumed to be met. σ is unknown, we will use one-sample t-test for mean.
Step 4. $t_{\text{st}}\,(\text{df}) = \frac{\bar{x}-\mu_0}{\frac{s}{\sqrt{n}}} = \frac{12.05-12.1}{\frac{0.085}{\sqrt{4}}} \approx -1.176, \quad df = n - 1 = 3.$
Step 5. p-value $= 2P(t\,(3) < -1.176) \approx 0.324$
Step 6. Since p-value $> \alpha$, H_0 is not rejected at 5% significance level.
Step 7. Thus, new don not have enough evidence to claim that the mean amount of juice dispensed is different from 12.1 fluid ounces.

As we can see from the boxplot 12 dots belong to values higher than 0.085. $P(X \geqslant 0.085) = \frac{12}{300} = 0.04$. Thus, the simulation shows that given $\sigma = 0.05$ only 4% of all samples of size 4 would produce sample standard deviation as high as $s = 0.085$ or even higher. This is the p-value for the test: H_0: $\sigma = 0.05$, $H_a : \sigma > 0.05$. Let's take $\alpha = 0.05$. Since p-value < 0.05, H_0 should be rejected at 5% significance level. Therefore, we think that the sample provides evidence that the standard deviation of the juice dispensed exceeds 0.05 fluid ounce.

Problem 8. AP 2008 №6
Administrators in a large school district wanted to determine whether students who attended a new magnet school for one year achieved greater improvement in science test performance than students who did not attend the magnet school. Knowing that more parents would want to enroll their children in the magnet school than there was space available for those children, the district administrators decided to conduct a lottery of all families who expressed interest in participating. In their data analysis,

the administrators would then compare the change in test scores of those children who were selected to attend the magnet school with the change in test scores of those who applied to attend the magnet school but who were not selected.

The tables below show the scores on the same science pretest and the same science posttest for 20 students. Of the 20 students, 8 were randomly selected from the magnet school and 12 were randomly selected from those who applied to attend the magnet school but who were not selected and then attended their original school.

(a) Perform a test to determine whether students who attend the magnet school demonstrate a significantly higher mean difference in test scores (Posttest – Pretest) than students who applied to attend the magnet school but who were not selected and then attended their original school.

(b) See Chapter 6.

(c) See Chapter 6.

(d) See Chapter 6.

Solution

Step 0. Let X_1 and X_2 be the increase in test scores (Posttest-Pretest) of children who attended Magnet school and Original school, correspondingly.

Step 1. $H_0 : \mu_1 - \mu_2 = 0$

$H_a : \mu_1 - \mu_2 > 0$

Let $\alpha = 0.05$

Step 2. Based on the table data:$\bar{X}_1 = 11.75$, $\bar{X}_2 = 3, s_1 = 9.407$, $s_2 = 3.977$, $n_1 = 8$, $n_2 = 12$.

Step 3. 1). Both sample are said to randomly selected from the corresponding population.

2). The sample sizes (< 30) are not large enough to ensure normality of $\bar{X}_1$ and$\bar{X}_2$ by CLT. The histograms below show distributions of X_1 and X_2. They reveal no obvious departure from normal distribution. So, we can assume normality of X_1 and X_2.

3). Since the district administrators had chosen students into Magnet school by a lottery, we can say that X_1 and X_2 are independent.

Step 4. $t_{\text{st}}\,(\text{df}) = \frac{\bar{X}_1 - \bar{X}_2 - 0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = \frac{(11.75-3)-0}{\sqrt{\frac{9.407^2}{8} + \frac{3.977^2}{12}}} \approx 2.487$, df ≈ 8.689 (taken from calculator).

Step 5. p-value $= P(t\,(8.7) > 2.487) \approx 0.018$

Step 6. Since p-value $< \alpha$, H_0 is rejected at 5% significance level[a].

Step 7. Thus, we have enough evidence that students who attend the magnet school demonstrate a significantly higher mean difference in test scores than students who applied to attend the magnet school but were not selected and attended their original school.

[a]Note, that if you choose $\alpha = 0.01$, p-value $> \alpha$, and H_0 would not be rejected at 1% significance level.

Problem 9. AP 2008 Form B №1
A certain state's education commissioner released a new report card for all the public schools in that state. This report card provides a new tool for comparing schools across the state. One of the key measures that can be computed from the report card is the student-to-teacher ratio, which is the number of students enrolled in a given school divided by the number of teachers at that school.

The data below give the student-to-teacher ratio at the 10 schools with the highest proportion of students meeting the state reading standards in the third grade and at the 10 schools with the lowest proportion of students meeting the state reading standards in the third grade.

Ratios in the 10 Schools with Highest Proportion of Students Meeting Standards
Ratios in the 10 Schools with Lowest Proportion of Students Meeting Standards

(a) See Chapter 6

(b) Any statistical test that is used to determine whether the mean student-to-teacher ratio is the same for the top 10 schools as it is for the bottom 10 schools would be inappropriate. Explain why in a few sentences.

Solution

(b) The statistical test comparing means should be based on the two samples taken from two independent populations. The two presented datasets are not samples, but populations of student-to-teacher ratios in the top 10 schools and the bottom 10 schools. There is no need to conduct a test to determine whether there is a difference in mean ratios in the two populations – it canbe calculated exactly.

Problem 10. AP 2008 Form B №4
A researcher wants to conduct a study to test whether listening to soothing music for 20 minutes helps to reduce diastolic blood pressure in patients with high blood pressure, compared to simply sitting quietly in a noise-free environment for 20 minutes. One hundred patients with high blood pressure at a large medical clinic are available to participate in this study.

(a)

(b) See Chapter 7.

(c) The null hypothesis for this study is that there is no difference in the mean reduction of diastolic blood pressure for the two treatments and the alternative hypothesis is that the mean reduction in diastolic blood pressure is greater for the music treatment. If the null hypothesis is rejected, the clinic will offer this music therapy as a free service to their patients with high blood pressure. Describe Type I and Type II errors and the consequences of each in the context of this study, and discuss which one you think is more serious.

Solution

(b) Type I error occurs when the true null hypothesis is mistakenly rejected. In the presented example it means that we would decide that the mean reduction in diastolic blood pressure is greater for the music treatment, although the music treatment is not effective. As a result the clinic will to no purpose offer this music therapy as a free service to their patients.

Type II error occurs when we fail to reject the false null hypothesis. Here it means to decide that there is no difference in the mean reduction of diastolic blood pressure for the two treatments, while music is an effective treatment for reducing blood pressure. As a result the clinic will not offer the music therapy, although doing this is an effective way to cure people.

We think that the type II error should be a more concern to the clinic[a].

[a]Alternatively, you can say that Type I error is more serious because it will cost the clinic money with no benefit. Both conclusions are admissible as long as they are accurately motivated.

Problem 11. AP 2007 №5

Researchers want to determine whether drivers are significantly more distracted while driving when using a cell phone than when talking to a passenger in the car. In a study involving 48 people, 24 people were randomly assigned to drive in a driving simulator while using a cell phone, while the remaining 24 were assigned to drive in the driving simulator while talking to a passenger in the simulator. Part of the driving simulation for both groups involved asking drivers to exit the freeway at a particular exit. In the study, 7 of the 24 cell phone users missed the exit and 2 of the 24 talking to a passenger missed the exit.

1. See Chapter 7
2. State the null and alternative hypotheses of interest to the researchers.
3. One test of significance that you might consider using to answer the researchers question is a two-sample z-test. State the conditions required for this test to be appropriate. The comment on whether each condition is met.
4. Using an advanced statistical method for small samples to test the hypothesis in part (b), the researchers report a p-value of 0.0683. Interpret, in everyday language, what this p-value measures in the context of this study *and* state what conclusion should be made based on this p-value.

Solution

2. Let p_1 be the population proportion of drivers who would miss the exit while using a cell phone and p_2 the population proportion of drivers who would miss the exit while talking to a passenger.

   $H_0 : p_1 - p_2 = 0$

$H_a : p_1 - p_2 > 0$

3. 1). Both samples should be randomly taken from populations. Nothing is stated about the way 48 people involved were chosen. However, we can assume that it was a random sample of drivers.

   2). The samples should be large enough to ensure normality of $\hat{p}_1$ and $\hat{p}_2$ by CLT. We have: $n_1\hat{p}_1 = 7 > 5$, $n_1(1-\hat{p}_1) = 17 > 5$, $n_2(1-\hat{p}_2) = 22 > 5$, but $n_2\hat{p}_2 = 2 < 5$. Then, this assumption is not satisfied.

   3). $\hat{p}_1$ and $\hat{p}_2$ should be independent. Since people were randomly assigned to the groups (to drive using a cell phone and talking to a passenger), we can say that the condition is satisfied.

4. p-value is the probability to get such an extreme samples or even more extreme given the null hypothesis is true.

   Here the samples produce the difference in proportions $\hat{p}_1 - \hat{p}_2$ equal to $\frac{7}{24} - \frac{2}{24} \approx 0.208$, which is quite far from zero, predicted by H_0. In the context of this study p-value means that if drivers were equally likely to be distracted while using a cell and when talking to a passenger, only 6.83% of all samples, randomly taken from the populations would produce an observation of at least 0.208.

   Let's take $\alpha = 0.1$. Since p-value0.1, H_0 is rejected at 10% significance level. Thus, we can conclude that using cell phone while driving is more distructive than talking.

Problem 12. AP 2006 Form B №6

Sunshine Farms wants to know whether there is a difference in consumer preference for the new juice products – Citrus Fresh and Tropical Taste. In an initial blind taste test, 8 randomly selected consumers were given unmarked samples of the two juices. The product that each consumer tasted first was randomly decided by the flip of a coin. After tasting the two juices each consumer was asked to choose which juice he or she preferred, and the results were recorded.

(a) Let p represent the population proportion of consumers who prefer Citrus Fresh. In terms of p, state the hypotheses that Sunshine Farms is interested in testing.

(b) One might consider using a one-proportion z-test to test the hypotheses in part (a). Explain why this would *not* be a reasonable procedure for this sample.

(c) See Chapter 2.

(d) testing the hypotheses in part (a), Sunshine Farms will conclude that there is a consumer preference if too many or too few individuals prefer Citrus Fresh. Based on your probabilities in part (c), is it possible for the significance level (probability of rejecting the null hypothesis when it is true) for this test to be exactly 0.05? Justify your answer.

(e) The preference data for 8 randomly selected consumers are given in the table below.

Based on these preferences and you previous work, test the hypotheses in part (a).

(f) Sunshine Farms plans to add one of these two juices – Citrus Fresh and Tropical Taste – to its production schedule. A follow-up study will be conducted to decide which of the two juices to produce. Make one recommendation for the follow-up study that would make it better than initial study. Provide a statistical justification for your recommendation in the context of the problem.

Solution

(c) If there is no difference in consumer preference for the new juice products, p=0.5. So,
$H_0 : p = 0.5$
$H_a : p \neq 0.5$

(d) The sample size is $n = 8$. Since $np_0 = n(1 - p_0) = 8 \cdot 0.5 = 4 < 5$, sample is not large enough to ensure normality of $\hat{p}$ (by CLT). We cannot use standard z-test procedure.

In part (c) we've shown that under assumption of no difference in consumer preference the distribution of the number of consumers in the sample who prefer Citrus Fresh X is as follows:

If the decision rule would be to reject H_0 when $X = 0$ or $X = 8$, $\alpha = 2 \cdot 0.00391 = 0.00782 < 0.05$. If we choose to reject H_0 when X$\leq$0 or X$\geq$ 7, $\alpha = 0.00782 + 2 \cdot 0.03215 = 0.07032 > 0.05$. For any other symmetric decision rules α would be higher. So, is it not possible for the significance level for this test to be exactly 0.05.

(e) Let's choose $\alpha = 0.1$. Based on the table data $X = 2$. Based on the probability distribution calculated in part (c) p-value $= 2P(X \leq 2) \approx 0.289$. Since p-value is higher than any reasonable α, H_0 cannot be rejected. Thus, there is no evidence of difference in consumer preference for the two juice products.

(f) We recommend to increase the sample size n. With larger sample size we would be able to state that $\hat{p} \sim N$ and use z-test. The variance of $\hat{p}$ is $\frac{p(1-p)}{n}$, it decreases with n, making estimated proportion $\hat{p}$ more accurate. The decision based on the test will also be done with higher degree of precision. For example, if we'd observe sample proportion of the same magnitude $\hat{p} = 0.2$ based on the sample of size 80, z-statistic would be $\frac{\hat{p}-p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} = \frac{0.25-0.5}{\sqrt{\frac{0.5(1-0.5)}{80}}} \approx -4.472$ and p-value$\approx 7.744 \cdot 10^{-6}$. In this situation we would be able to reject H_0.

Problem 13. AP 2005 №4
Some boxes of a certain brand of breakfast cereal include a voucher for a free video rental inside the box. The company that makes the cereal claims that a voucher can be found in 20 percent of the boxes. However, based on their experiences eating this cereal at home, a group of students believes that the proportion of boxes with voucher is less than 0.2. This group of students purchased 65 boxes of the cereal to investigate the company's claim. The students found a total of 11 vouchers for free video rentals in the 65 boxes.

Suppose it is reasonable to assume that the 65 boxes purchased by the students are a random sample of all boxes of this cereal. Based on this example, is there support for the students' belief that the proportion of boxes with vouchers is less than 0.2? Provide statistical evidence to support your answer.

Solution
(a) Step 0. Let p be the proportion of boxes with a voucher.
Step 1. H_0: $p = 0.2$
H_a: $p < 0.2$
Let's choose $\alpha = 0.05$
Step 2. $\hat{p} = \frac{11}{65} \approx 0.169$, $n = 65$.
Step 3. The assumption of random sample is stated to be reasonable.
$np_0 = 65 \cdot 0.2 = 13 > 5$, $n(1 - p_0) = 65 \cdot 0.8 = 52 > 5$. So, sample is large enough to ensure normality of $\hat{p}$ (by CLT). We will use one-sample z-test for proportion.
Step 4. $z_{\text{st}} = z_{\text{st}} = \frac{\hat{p} - p_0}{\sqrt{\frac{p(1-p_0)}{n}}} = \frac{0.169 - 0.2}{\sqrt{\frac{0.2(1-0.2)}{65}}} \approx -0.620$
Step 5. p-value$= P(z < -0.62) \approx 0.268$
Step 6. Since p-value $> \alpha$, H_0 is not rejected at 5% significance level.
Step 7. Thus, there is no support for the students' belief that the proportion of boxes with vouchers is less than 0.2.

Problem 14. 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) See Chapter 10.

(b) Because the 95 percent confidence interval includes 20, the regulatory agency is not convinced that the new drug is better than the current best-seller. The pharmaceutical company tested the following hypothesis.

H_0: $\mu = 20$ versus H_a: $\mu > 20$

Where μ represents the population mean reduction in cholesterol level for the new drug.

The test procedure resulted in a t-value of 1.89 and a p-value of 0.033. Because the p-value was less than 0.05, the company believes that there is convincing evidence that the mean reduction in cholesterol level for the new drug is more than 20. Explain why the confidence interval and the hypothesis test led to different conclusions.

(c) See Chapter 10.

(d) See Chapter 10.

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

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

(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

(b) in part (a) the two-sided 95% confidence interval $\mu \in [19.737; 28.263]$ was calculated, which gave no evidence for the $\mu > 20$ hypothesis. Contrary, here the test with a one-sided alternative leads to the decision in favor of $\mu > 20$ hypothesis at 5% significance level. The difference is explained by the one-sided/two-sided option. The two-sided test with significance level α would produce the same decision as the two-sided interval with confidence level $1-\alpha$ did. The two-sided test based on the same data provides p-value $= 2P(t(49) > 1.89) = 2 \cdot 0.033 = 0.066 > 0.05$. This leads to the same conclusion of no evidence for the $\mu > 20$ hypothesis as the interval provides. There is no contradiction.

Equivalently, one-sided test with (at level α) produces the same decision as the one-sided $(1 - \alpha)$ interval. You can check it in part (e).

Problem 15. AP 2003 №1

Since Hill Valley High School eliminates the use of bells between classes, teachers had noticed that more students seem to be arriving to class a few minutes late. One teacher decided to collect data to determine whether the students' and teachers' watches are displaying the correct time. At exactly 12:00 noon the teacher asked 9 randomly selected students and 9 randomly selected teachers to record the time on their watches to the nearest half minute. The ordered data showing minutes after 12:00 as positive values as minutes before 12:00 as negative values are shown in the table below.

Students	-4.5	-3.0	-0.5	0	0	0.5	0.5	1.5	5.0
Teachers	-2.0	-1.5	-1.5	-1.0	-1.0	-0.5	0	0	0.5

(a) See Chapter 6.

(b) See Chapter 6.

(c) The teacher wants to know whether individual student's watches tend to be set correctly. She proposes to test H_0: $\mu = 0$ versus H_a: $\mu \neq 0$, where μ represents the mean amount by which all student watches differ from the correct time. Is this an appropriate pair of hypotheses to test to answer the teacher's question? Explain why or why not. *Do not carry out the test.*

Solution

(c) This is not an appropriate way to check whether student's watches tend to be set correctly. The proposed test will check whether mean deviation is zero. However, since some deviations are positive and some are negative the sample mean will tend to reduce to zero. Thus, even with large deviations, were positive and negative balanced, we may fail to reject H_0: $\mu = 0$. So, this test does not answer the teacher's question.

Practice AP problems

Problem 1. AP 2012 №4
A survey organization conducted telephone interviews in December 2008 in which 1,009 randomly selected adults in the United States responded to the following question.

At the present time, do you think television commercials are an effective way to promote a new product?

Of the 1,009 adults surveyed, 676 responded "yes." In December 2007, 622 of 1,020 randomly selected adults in the United States had responded "yes" to the same question. Do the data provide convincing evidence that the proportion of adults in the United States who would respond "yes" to the question changed from December 2007 to December 2008?

Problem 2. AP 2010 №5
A large pet store buys the identical species of adult tropical fish from two different suppliers—Buy-Rite Pets and Fish Friends. Several of the managers at the pet store suspect that the lengths of the fish from Fish Friends are consistently greater than the lengths of the fish from Buy-Rite Pets. Random samples of 8 adult fish of the species from Buy-Rite Pets and 10 adult fish of the same species from Fish Friends were selected and the lengths of the fish, in inches, were recorded, as shown in the table below.

	Length of Fish	Mean	Standard Deviation
Buy-Rite Pets ($n_B = 8$)	3.4 2.7 3.3 4.1 3.5 3.4 3.0 3.8	3.40	0.434
Fish Friends ($n_F = 10$)	3.3 2.9 4.2 3.1 4.2 4.0 3.4 3.2 3.7 2.6	3.46	0.550

Do the data provide convincing evidence that the mean length of the adult fish of the species from Fish Friends is greater than the mean length of the adult fish of the same species from Buy-Rite Pets?

Problem 3. AP 2009 №6

A consumer organization was concerned that an automobile manufacturer was misleading customers by overstating the average fuel efficiency (measured in miles per gallon, or mpg) of a particular car model. The model was advertised to get 27 mpg. To investigate, researchers selected a random sample of 10 cars of that model. Each car was then randomly assigned a different driver. Each car was driven for 5,000 miles, and the total fuel consumption was used to compute mpg for that car.

(a) Define the parameter of interest and state the null and alternative hypotheses the consumer organization is interested in testing.

(b) See Chapter 6

(c) See Chapter 6

Problem 4. AP 2009 Form B №3

A French study was conducted in the 1990s to compare the effectiveness of using an instrument called acardiopump with the effectiveness of using traditional cardiopulmonary resuscitation (CPR) in saving lives of heart attack victims. Heart attack patients in participating cities were treated with either a cardiopump or CPR, depending on whether the individual's heart attack occurred on an even-numbered or an odd-numbered day of the month. Before the start of the study, a coin was tossed to determine which treatment, a cardiopump or CPR, was given on the even-numbered days. The other treatment was given on the odd-numbered days. In total,

754 patients were treated with a cardiopump, and 37 survived at least one year; while 746 patients were treated with CPR, and 15 survived at least one year.

(a) The conditions for inference are satisfied in the study. State the conditions and indicate how they are satisfied.

(b) Perform a statistical test to determine whether the survival rate for patients treated with a cardiopump is significantly higher than the survival rate for patients treated with CPR.

Problem 5. AP 2008 Form B №6

The nerves that supply sensation to the front portion of a person's foot run between the long bones of the foot.

Tight-fitting shoes can squeeze these nerves between the bones, causing pain when the nerves swell. This condition is called Morton's neuroma. Because most people have a dominant foot, muscular development is not the same in both feet. People who have Morton's neuroma may have the condition in only one foot or they may have it in both feet.

Investigators selected a random sample of 12 adult female patients with Morton's neuroma to study this disease further. The data below are measurements of nerve swelling as recorded by a physician. A value of 1.0 is considered "normal," and 2.0 is considered extreme swelling. The population distribution of the swelling measurements is approximately normal for adult females who have Morton's neuroma.

Dominant Foot	Swelling in Dominant Foot	Swelling in Nondominant Foot	Foot with Neuroma
Left	1.40	1.10	Left
Left	1.55	1.25	Left
Left	1.65	1.20	Left
Left	1.55	1.40	Both
Left	1.70	1.40	Left
Left	1.85	1.50	Both
Right	1.45	1.20	Right
Right	1.65	1.30	Right
Right	1.60	1.40	Right
Right	1.70	1.45	Both
Right	1.85	1.45	Both
Right	1.75	1.60	Both

(a) See Chapter 13

(b) See Chapter 13

(c) Can you conclude that there is a difference between the mean swelling in the dominant foot and the mean swelling in the nondominant foot for adult females who have Morton's neuroma in at least one foot? Give a statistical justification to support your answer.

(d) See Chapter 6

Problem 6. AP 2007 №4
Investigators at the U.S. Department of Agriculture wished to compare methods of determining the level of E. coli bacteria contamination in beef. Two different methods (A and B) of determining the level of contamination were used on each of ten randomly selected specimens of a certain type of beef. The data obtained, in millimicrobes/liter of ground beef, for each of the methods are shown in the table below.

		Specimen									
		1	2	3	4	5	6	7	8	9	10
Method	A	22.7	23.6	24.0	27.1	27.4	27.8	34.4	35.2	40.4	46.8
	B	23.0	23.1	23.7	26.5	26.6	27.1	33.2	35.0	40.5	47.8

Is there a significant difference in the mean amount of E. coli bacteria detected by the two methods for this type of beef? Provide a statistical justification to support your answer.

Problem 7. AP 2007 Form B №5
A serum cholesterol level above 250 milligrams per deciliter (mg/dl) of blood is a risk factor for cardiovascular disease in humans. At a medical center in St. Louis, a study to test the effectiveness of a new cholesterol-lowering drug was conducted. One hundred people with cholesterol levels between 250 mg/dl and 300 mg/dl were available for this study. Fifty people were assigned at random to each of two treatment groups. One group received the standard cholesterol-lowering medication and the other group received the new drug. After taking the drug for three weeks, the 50 subjects who received the standard treatment had a mean decrease in cholesterol level of 10 mg/dl with a standard deviation of 8 mg/dl, and the 50 subjects who received the new drug had a mean decrease of 18 mg/dl with a standard deviation of 12 mg/dl.

Does the new drug appear to be more effective than the standard treatment in lowering mean cholesterol level? Give appropriate statistical evidence to support your conclusion.

Problem 8. AP 2007 Form B №6 (a)
Scientists interested in preserving natural habitats and minimizing the possible extinction of certain bird species conducted a study to determine if it is better for conservation groups to purchase a few large nature preserves or many small preserves in order to meet these goals.

The scientists studied 13 randomly selected islands of different sizes to determine the risk of extinction for bird species. Islands are thought to be a good imitation of

what would happen in a nature preserve because of their isolation. If a species lived on only one island, it was considered to be at risk. Scientists have determined that whether or not one species becomes extinct is independent of whether or not another species becomes extinct.

In 1990 scientists counted the number of at-risk species on each of the selected islands. They returned to each of these islands in the year 2000 to see whether the species still existed on the islands. Species that were present in 1990 but absent in 2000 were considered extinct. Data collected by the scientists are given in the table below.

(a) One scientist involved in the study believes that large islands (those with areas greater than 25 square kilometers) are more effective than small islands (those with areas of no more than 25 square kilometers) for protecting at-risk species. The scientist noted that for this study, a total of 19 of the 208 species on the large islands became extinct, whereas a total of 66 of the 299 species on the small islands became extinct.

Assume that the probability of extinction is the same for all at-risk species on large islands and the same for all at-risk species on small islands. Do these data support the scientist's belief? Give appropriate statistical justification for your answer.

Island	Area	Species at risk in 1990	Species Extinct by 2000	Proportion Extinct
1	46	75	8	0,11
2	36	67	3	0,04
3	31	66	8	0,12
4	9	51	8	0,16
5	5	28	5	0,18
6	5	20	6	0,30
7	4	43	10	0,23
8	4	31	5	0,16
9	3	28	7	0,25
10	2	32	8	0,25
11	1	30	8	0,27
12	1	20	4	0,20
13	1	16	5	0,21

Problem 9. AP 2006 Form B №4

The developers of a training program designed to improve manual dexterity claim that people who complete the 6-week program will increase their manual dexterity. A random sample of 12 people enrolled in the training program was selected. A measure of each person's dexterity on a scale from 1 (lowest) to 9 (highest) was recorded just before start of and just after the completion of 6-week program. The data are shown in the table below. Can one conclude that the mean manual dexterity for people who have completed the 6-week training program has significantly increased? Support your conclusion with appropriate statistical evidence.

Problem 10. AP 2003 №2
When a law firm represents a group of people in a class action lawsuit and wins that lawsuit, the firm receives a percentage of the group's monetary settlement. The settlement amount is based on the total number of people in the group – the larger the group and the larger the settlement, the more money the firm will

A law firm is trying to decide whether to represent car owners in a class action lawsuit against the manufacturer of a certain make and model for a particular defect. If 5 percent or less of the cars of this make and model have the defect, the firm will not recover its expenses. Therefore, the firm will handle the lawsuit only if it is convinced that more than 5 percent of cars of this make and model have the defect. The firm plans to take a random sample of 1,000 people who bought this car and ask them if they experienced this defect in their cars.

(b) In the context of this situation, describe Type I and Type II errors and describe the consequences of each of these for the law firm.

Problem 11. AP 2003 Form B №3 (b)
A study was conducted to determine if taking vitamin C reduces the occurrence of the flu. The study was conducted using 808 students volunteers who did not take a flu shot. The subjects were randomly assigned to one of two group: a treatment group who received 1,000 milligrams of vitamin C daily or a control group who received a placebo flavored to taste like Vitamin C treatment. All participants were monitored to ensure that they adhered to their assigned treatment on a daily basis throughout the period of the study. The physician did not know which treatment each subject received. The results of the study are shown in the table below.

	Flu	No Flu	Total
Placebo	331	74	405
Vitamin C	302	101	403
Total	633	175	808

(b) Based on this study, a health expert claims that there is evidence to suggest that vitamin C reduces the occurrence of the flu in the population of students who would volunteer for such a study. State the name of a test and the null and alternative hypothesis that the health expert could have used to support this claim. *Do not* carry out the test.

Problem 12. AP 2003 Form B №4 (c)
There have been many studies recently concerning coffee drinking and cholesterol level. While it is known that several coffee-bean components can elevate blood cholesterol level, it is through that a new type of paper coffee filter may reduce the presence of some of these components in coffee.

The effect of the new filter on cholesterol level will be studied over a 10-week period using 300 nonsmokers who each drink 4 cups of caffeinated coffee per day. Each of these 300 participants will be assigned to one of two groups: the experimental group,

who will only drink coffee that has been made with the new filter, or the control group, who will only drink coffee, that has been made with the new filter, or the control group, who will only drink coffee, that has been made with the standard filter. Each participant's cholesterol level will be measured at the beginning and at the end of the study.

(c) Which test would you conduct to determine whether the change in cholesterol level would be greater if people used the new filter rather than using the standard filter?

Problem 13. AP 2002 №5

Sleep researches know that some people are early birds (E), preferring to go to bed by 10 p.m. and arise by 7 a.m., while others are night owls (N), preferring to go to bed after 11 p.m. and arise after 8 a.m. A study was done to compare dream recall for early birds and night owls. One hundred people of each of the two types were selected at random and asked to record their dreams for one week. Some of the results are presented below.

	Number of Dreams Recalled During the Week			Proportion Who Recalled	
Group	Mean	Median	Standard Deviation	No dreams	5 or more dreams
Early birds	7.26	6.0	6.94	0.24	0.55
Night owls	9.55	9.5	5.88	0.11	0.69

(a) The researches believe that night owls may have better dream recall than do early birds.

One parameter of interest to the researches is the mean number of dreams recalled per week with μ_E representing this mean for early birds and μ_N representing this mean for night owls. The appropriate hypothesis would then be H_0: $\mu_E - \mu_N = 0$ and H_A: $\mu_E - \mu_N < 0$. State two other pairs of hypotheses that might be used to test the researches' belief. Be sure to define the parameter of interest in each case.

(b) Use the data provided to carry out a test of the hypotheses about the mean number of dreams recalled per week given in the statement of part (a). Do the data support the researches' belief?

Problem 14. AP 2002 Form B №6

In September 1990, each student in a random sample of 200 biology majors at a large university was asked how many lab classes he or she was enrolled in. The sample results are shown below.

Number of lab classes	Number of students
0	28
1	62
2	58
3	28
4	16
5	8
(Total)	200

$$\bar{x} = 1.83,\ s = 1.29$$

To determine whether the distribution has changed over the past 10 years, a similar survey was conducted in September 2000 by selecting a random sample of 200 biology majors. Results from the year 2000 sample are shown below.

Number of lab classes	Number of students
0	20
1	72
2	60
3	10
4	26
5	12
(Total)	200

$$\bar{x} = 1.93,\ s = 1.37$$

(a) Do the data provide evidence that the mean number of lab classes taken by biology majors in September 2000 was different from the mean number of lab classes taken in 1990? Perform an appropriate statistical test using $\alpha = 0.10$ to answer this question.

(b) See chapter 12.

(c) See chapter 12.

Problem 15. AP 2001 №5

A growing number of employers are trying to hold down the costs that they pay for medical insurance for their employees. As part of this effort, many medical insurance companies are now requiring clients to use generic brand medicines when filling prescriptions. An independent consumer advocacy group wanted to determine if there was a difference, in milligrams, in the amount of active ingredient between a certain "name" brand drug and its generic counterpart. Pharmacies may store drugs under different conditions. Therefore the consumer group randomly selected ten different pharmacies in a large city and filled two prescriptions in each of these pharmacies, one for the "name" brand and the other for the generic brand of the drug. The consumer group's laboratory then tested a randomly selected pill from each prescription to determine the amount of active ingredient in the pill. The results are given in the following table.

Pharmacy	1	2	3	4	5	6	7	8	9	10
Name brand	245	244	240	250	243	246	246	246	247	250
Generic brand	246	240	235	237	243	239	241	238	238	234

Based on these results, what should the consumer group's laboratory report about the difference in the active ingredient in the two brands of pills? Give appropriate statistical evidence to support your response.

Problem 16. AP 2000, №4

Baby walkers are seats hanging from frames that allow babies to sit upright with their legs dangling and feet touching the floor. Walkers have wheels on their legs that allow the infant to propel the walker around the house long before he or she can walk or even crawl. Typically, babies use walkers between the ages of 4 months and 11 months.

Because most walkers have tray tables in front that block babies' views of their feet, child psychologists have begun to question whether walkers affect infants' cognitive development. One study compared mental skills of a random sample of those who used walkers with a random sample of those who never used walkers. Mental skill scores averaged 113 for 54 babies who used walkers (standard deviation of 12) and 123 for 55 babies who did not use walkers (standard deviation of 15).

(a) Is there evidence that the mean mental skill score of babies who use walkers is different from the mean mental skill score of babies who do not use walkers? Explain your answer.

(b) Suppose that a study using this design found a statistically significant result. Would it be reasonable to conclude that using a walker causes a change in mean mental skill score? Explain your answer.

Problem 17. AP 1999 №6 (a-b)

Researchers want to see whether training increases the capability of people to correctly predict outcomes of coin tosses. Each of twenty people is asked to predict the outcome (heads or tails) of 100 independent tosses of a fair coin. After training, they are retested with a new set of 100 tosses. (All 40 sets of 100 tosses are independently generated.) Since the coin is fair, the probability of a correct guess by chance is 0.5 on each toss. The numbers correct for each of the 20 people were as follows.

Score Before Training (number correct)	Score After Training (number correct)
46	61
48	62
50	53
54	46
54	50
54	52
54	53
54	59
54	60
54	61
55	55
56	59
57	55
58	50
58	56
61	58
61	64
63	57
64	61
65	54
Sum 1,120	Sum 1,126

To answer the following questions, you may want to enter these data into your calculator. As a check that you have entered the data correctly, the sum of the first column is 1,120 and the sum of the second column is 1,126.

(a) Do the data suggest that after training people can correctly predict coin toss outcomes better than the 50 percent expected by chance guessing alone?

Give appropriate statistical evidence to support your conclusion.

(b) Does the statistical test that you completed in part (a) provide evidence that this training is effective in improving a person's ability to predict coin toss outcomes?

If yes, justify your answer. If no, conduct an appropriate analysis that would allow you to determine whether or not the training is effective.

Problem 18. AP 1998 №5

A large university provides housing for 10 percent of its graduate students to live on campus.

The university's housing office thinks that the percentage of graduate students looking for housing on campus may be more than 10 percent. The housing office decides to survey a random sample of graduate students, and 62 of the 481 respondents say that they are looking for housing on campus.

(a) On the basis of the survey data, would you recommend that the housing office consider increasing the amount of housing on campus available to graduate students? Give appropriate evidence to support your recommendation.

(b) In addition to the 481 graduate students who responded to the survey, there were 19 who did not respond. If these 19 had responded, is it possible that your recommendation would have changed? Explain.

Problem 19. AP 2019 №4
Tumbleweed, commonly found in the western United States, is the dried structure of certain plants that are blown by the wind. Kochia, a type of plant that turns into tumbleweed at the end of the summer, is a problem for farmers because it takes nutrients away from soil that would otherwise go to more beneficial plants. Scientists are concerned that kochia plants are becoming resistant to the most commonly used herbicide, glyphosate. In 2014, 19.7 percent of 61 randomly selected kochia plants were resistant to glyphosate. In 2017, 38.5 percent of 52 randomly selected kochia plants were resistant to glyphosate. Do the data provide convincing statistical evidence, at the level of $\alpha = 0.05$ that there has been an increase in the proportion of all kochia plants that are resistant to glyphosate?

Answers to practice problems

Problem 1. $z_{\text{st}} = -2.82$. p-value = 0.0048. The proportion of all adults in the United States who would answer "yes" to the question about the effectiveness of television commercials changed from December 2007 to December 2008.

Problem 2. $t_{\text{st}}(15.99) = -0.259$, p-value = 0.3996. The sample data do not provide convincing evidence to conclude that the mean length of the adult fish of the species from Fish Friends is greater than the mean length of the adult fish of the same species from Buy-Rite Pets.

Problem 3. (a) H_0: $m = 27$, H_a: $m < 27$.

Problem 4. (a) The conditions required for a two-sample z test of equal proportions for an experiment are: 1. Random assignment of treatments to subjects 2. Sufficiently large sample sizes

(b) $z_{\text{st}} = 3.006$, p-value = 0.0011. We have strong evidence to support the conclusion that the proportion of patients who survive when treated with the cardiopump is higher than the proportion of patients who survive when treated with CPR.

Problem 5. (c) $t_{\text{st}}(11) = 10.68$, p-value ≈ 0. There is convincing evidence that the mean swelling is different for dominant and nondominant feet for women with MN.

Problem 6. $t_{\text{st}}(9) = 1.46$, p-value = 0.179. There is no significant difference in the mean amount of E. coli bacteria detected by the two methods.

Problem 7. $t_{\text{st}}(df) = -3.92$, $df = 85$, p-value = 0.000088 OR $df = 49$, p-value = 0.00014 OR $zst = -3.92$, p-value = 0.000044 OR t_{st} for pooled t-test is -3.92, $df = 98$, p-value = 0.000081. There is convincing evidence that the mean cholesterol reduction is greater for the new drug.

Problem 8. $z_{\text{st}} = -3.84$, p-value = 0.00006. There is sufficient evidence that the proportion is smaller for large islands.

Problem 9. $t_{\text{st}}(11) = 3.54$, p-value = 0.002. People who completed the program have significantly increased manual dexterity.

Problem 10. Type I error: The law firm believes that the proportion of cars that have the defect is greater than 0.05, when in fact it is not. The firm will not recover its expences, resulting in a loss to the firm.

Type II error: The law firm is not convinced that the proportion of cars that have the defect is greater than 0.05, when in fact it is. The firm will miss an opportunity to make money on this case.

Problem 11. H_0: $p_r - p_c = 0$, H_A: $p_r - p_c < 0$.

Problem 12. The 2-sample t-test for means or mean differences would be used (or z-test for means).

Problem 13. (a) Null hypothesis of equal proportions of people recalling dreams among early birds and night owls with one-sided alternative.

(b) $t_{\text{st}}(192) = -2.52$, p-value = 0.006. It also Ok to use pooled samples t-test of z-test here. There is convincing evidence that the mean number of dreams night owls recall is higher.

Problem 14. $t_{\text{st}}(396) = -0.751$, p-value = 0.453. There is no convincing evidence that the mean number of lab classes in 2000 is different than it was in 1990.

Problem 15. $t_{\text{st}}(9) = -3.96$, p-value = 0.00332. There is evidence that the mean

amount of active ingredient is not the same for the name brand and generic drugs.

Problem 16. (a) $t_{\text{st}}(103) = -3.85$, p-value = 0.0002. It is also Ok to use pooled t-test or 2-sample z-test. There is convincing evidence that the mean mental score of babies who used walkers is different from the mean score for babies who did not use walkers.

(b) No, this was an observational study, so, causal inferences cannot be inferred from it.

Problem 17. (a) $t_{\text{st}}(19) = 5.963$, p-value = 0.0000049. z-test with normal approximation to proportion is also Ok. There is convincing evidence that the mean number of correct responses after the training is higher than 50, the value expected by chance alone.

(b) No, it does not, since this analysis does not compare before and after scores. Paired samples t-test: $t_{\text{st}}(19) = 0.1962$, p-value = 0.4233. No evidence of training effect.

Problem 18. (a) 2-sample z-test for proportion. $z_{\text{st}} = 2.113$, p-value = 0.017. There is evidence that true proportion is higher than 10% (at 5% significance level).

(b) Of course, is all 19 need housing, the recommendation will not change. If all 19 do not need housing: $zst = 1.789$, p-value = 0.037. At 5% significance level, proportion is still decided to be higher than 10%.

Problem 19. $z_{\text{st}} = 2.21$, p-value = 0.0135. There is convincing statistical evidence to conclude that the proportion of resistant plants in the 2017 population of kochia plants is greater than the proportion of resistant plants in the 2014 population of kochia plants.