Chapter 3

Confidence intervals and hypothesis testing

Exercises

Contents

3.1 Concrete items

|||| Exercise 3.1 Concrete items

A construction company receives concrete items for a construction. The length of the items are assumed reasonably normally distributed. The following requirements for the length of the elements are made

$$\mu = 3000 \text{ mm}.$$

The company samples 9 items from a delivery which are then measured for control. The following measurements (in mm) are found:

3003	3005	2997	3006	2999	2998	3007	3005	3001

a) Compute the following three statistics: the sample mean, the sample standard deviation and the standard error of the mean, and what are the interpretations of these statistics?

|||| Solution

From the data we get the sample mean and sample standard deviation

$$\bar{x} = 3002.33 \text{ mm} \quad \text{and} \quad s = 3.708 \text{ mm}.$$

From the definition of the standard error of the mean:

$$SE_{\bar{x}} = \frac{s}{\sqrt{n}} = \frac{3.708}{\sqrt{9}} = 1.236.$$

Interpretations:

$\bar{x} = 3002.33$ The best estimate of the true mean length of such concrete items.

$s = 3.708$ The population standard deviation is estimated at 3.7 mm, meaning most items will be within approximately ± 7.4 mm of the mean.

$SE_{\bar{x}} = 1.236$ The estimated standard deviation of the sampling distribution of the sample mean. On average, the sample mean differs from the true population mean by 1.24 mm.

```
x = np.array([3003, 3005, 2997, 3006, 2999, 2998, 3007,
   3005, 3001])
print(x.mean(), x.std(ddof=1), x.std(ddof=1)/np.sqrt(len(x))
   )
```

b) In a construction process, 5 concrete items are joined together to a single construction with a total length equal to the sum of the 5 concrete items. It is very important that the length of this new construction is within 15 m plus/minus 1 cm. How often will it happen that such a construction will be more than 1 cm away from the 15 m target (assume that the population mean concrete item length is $\mu = 3000$ mm and the population standard deviation is $\sigma = 3$)?

|||| Solution

Let $Y = \sum_{i=1}^{5} X_i$ denote the length of the joined construction. Using the rules for linear combinations of independent normals:

$$\mathrm{E}(Y) = 5 \times 3000 = 15000,$$
$$\mathrm{V}(Y) = 5 \times 3^2 = 45, \quad Y \sim N(15000,\ 45).$$

$$P(|Y - 15000| > 10) = 2 \cdot P\left(Z > \frac{10}{\sqrt{45}}\right) = 2 \cdot P(Z > 1.491) = 0.136.$$

About 13–14 out of 100 joined constructions will be more than 1 cm off target.

```
import scipy.stats as stats
print(2*(1 - stats.norm.cdf(10, 0, np.sqrt(45))))
```

c) Find the 95% confidence interval for the mean μ.
(Quantile: $t_{0.975,8} = 2.306$)

|||| Solution

Using the t-based CI with $\bar{x} = 3002.33$, $s = 3.708$, $n = 9$ and $t_{0.975,8} = 2.306$:

$$3002.33 \pm 2.306 \cdot \frac{3.708}{\sqrt{9}} \iff [2999.5;\ 3005.2].$$

```
ci = stats.t.interval(0.95, df=8, loc=3002.33, scale=3.708/
    np.sqrt(9))
print(ci)
```

d) Find the 99% confidence interval for μ. Compare with the 95% one from above and explain why it is smaller/larger.
(Quantile: $t_{0.995,8} = 3.355$)

|||| Solution

$$3002.33 \pm 3.355 \cdot \frac{3.708}{\sqrt{9}} \iff [2998.2;\ 3006.5].$$

The 99% interval is wider: wanting to be more confident of capturing the true mean μ requires a broader interval.

e) Find the 95% confidence intervals for the variance σ^2 and the standard deviation σ.
(Quantiles: $\chi^2_{0.975,8} = 17.535$, $\chi^2_{0.025,8} = 2.180$)

|||| Solution

Using the χ^2-based formula with $(n-1)s^2 = 8 \times 3.708^2 = 8 \times 13.75 = 110.0$:

$$\left[\frac{8 \times 13.75}{17.535};\ \frac{8 \times 13.75}{2.180}\right] = [6.27;\ 50.46].$$

For the standard deviation:

$$[\sqrt{6.27};\ \sqrt{50.46}] = [2.50;\ 7.10].$$

f) Find the 99% confidence intervals for the variance σ^2 and the standard deviation σ.
(Quantiles: $\chi^2_{0.995,8} = 21.955$, $\chi^2_{0.005,8} = 1.344$)

|||| Solution

$$\left[\frac{8 \times 13.75}{21.955}; \frac{8 \times 13.75}{1.344}\right] = [5.01;\ 81.82].$$

For the standard deviation:

$$[\sqrt{5.01};\ \sqrt{81.82}] = [2.24;\ 9.05].$$

Exam-style questions

g) A machine shop samples $n = 16$ steel rods and measures their diameters, finding $\bar{x} = 25.0$ mm and $s = 0.40$ mm. What is the standard error of the mean, and which statement correctly interprets it?

1) $SE_{\bar{x}} = 0.10$ mm; it estimates how much the sample mean would typically vary from sample to sample of the same size, *not* how much individual rods vary.
2) $SE_{\bar{x}} = 0.40$ mm; it is the same as the sample standard deviation.
3) $SE_{\bar{x}} = 1.60$ mm; it estimates the typical spread of individual rod diameters.
4) $SE_{\bar{x}} = 0.025$ mm, computed as s/n.
5) $SE_{\bar{x}} = 0.10$ mm; it estimates how much individual rod diameters vary around the mean.

|||| Solution

Correct answer: **1)**. $SE_{\bar{x}} = s/\sqrt{n} = 0.40/\sqrt{16} = 0.10$ mm. As in the first question of this exercise, the standard error describes the sampling variability of $\bar{x}$ itself, not the spread of individual observations (which is instead $s = 0.40$ mm) — this is exactly the distinction option 5) gets wrong despite using the correct number. Option 3) inflates rather than shrinks s. Option 4) divides by n instead of $\sqrt{n}$.

h) 8 metal rods, each with length $X_i \sim N(500, 2^2)$ mm independently, are welded end to end. The total length must be within ± 3 mm of the 4000 mm target. What is the probability that the welded assembly misses this tolerance?

1) ≈ 0.596
2) ≈ 0.298
3) ≈ 0.134
4) ≈ 0.404
5) ≈ 0.851

|||| Solution

Correct answer: **1)**. Let $Y = \sum_{i=1}^{8} X_i \sim N(4000, 8 \times 2^2) = N(4000, 32)$, so $\sigma_Y = \sqrt{32} = 5.657$:

$$P(|Y - 4000| > 3) = 2\,P\left(Z > \frac{3}{5.657}\right) = 2\,P(Z > 0.530) \approx 0.596.$$

Option 2) forgets to double the tail probability. Option 3) mistakenly uses the single-rod standard deviation $\sigma = 2$ directly instead of scaling up for the sum of 8 rods. Option 4) is the complement of the correct answer. Option 5) mistakenly scales the standard deviation linearly with n (using $\sigma = 8 \times 2 = 16$) instead of by $\sqrt{n}$.

```
print(2*(1 - stats.norm.cdf(3, 0, np.sqrt(8*2**2))))
```

3.2 Aluminum profile

|||| Exercise 3.2 Aluminum profile

The length of an aluminum profile is checked by taking a sample of 16 items whose lengths are measured. The measurement results from this sample are listed below, all measurements are in mm:

180.02	180.00	180.01	179.97	179.92	180.05	179.94	180.10
180.24	180.12	180.13	180.22	179.96	180.10	179.96	180.06

From the data: $\bar{x} = 180.05$ mm and $s = 0.0959$ mm.

It can be assumed that the sample comes from a normally distributed population.

a) Compute the 90%-confidence interval for μ.
   Use Python to get the needed quantile from a t-distribution ($t_{1-\alpha/2}$).

|||| Solution

We use Method 3.9.

For the calculation we need $t_{1-\alpha/2}$ from a t-distribution with 15 degrees of freedom (since the number of observations is: $n = 16$, the degrees of freedom are calculated as: $\nu = n - 1 = 15$).

Since we are asked to compute the 90%-confidence interval we need the $t_{0.95}$-quantile (you can convince yourself of this by drawing the pdf of a t-distribution and sketch the interval (centered about the mean) corresponding to 90% of the total probability mass (area under the pdf-curve) - this will leave 5% of the total probability mass on either side of the interval).

We use Python to find the value of the $t_{0.95}$-quantile:

```
print(stats.t.ppf(0.95, df=15))
1.753
```

Using $\bar{x} = 180.05$, $s = 0.0959$, $n = 16$ and $t_{0.95,15} = 1.753$, we now compute the confidence interval:

$$180.05 \pm 1.753 \cdot \frac{0.0959}{\sqrt{16}} = 180.05 \pm 0.042 = [180.01;\ 180.09].$$

b) Compute the 99%-confidence interval for σ.
Use Python to get the needed quantiles from a χ^2-distribution ($\chi^2_{\alpha/2}$ and $\chi^2_{1-\alpha/2}$).

|||| Solution

We use Method 3.19.

For the calculation we need $\chi^2_{\alpha/2}$ and $\chi^2_{1-\alpha/2}$ from a χ^2-distribution with 15 degrees of freedom ($\nu = n - 1 = 15$).

Since we are asked to compute the 99%-confidence interval we need the $\chi^2_{0.005}$-quantile and the $\chi^2_{0.995}$-quantile (you can convince yourself of this by drawing the pdf of a χ^2-distribution and sketch the interval corresponding to 99% of the total probability mass (area under the pdf-curve), leaving 0.5% of the total probability mass on the left side of the interval and 0.5% of the total probability mass on the right side of the interval).

We use Python to find the values of the quantiles:

```
print(stats.chi2.ppf(0.005, df=15))
4.601

print(stats.chi2.ppf(0.995, df=15))
32.801
```

Using Method 3.19 with $(n-1)s^2 = 15 \times 0.0959^2 = 0.1380$, we get a confidence interval for the variance σ^2:

$$\left[\frac{0.1380}{32.801};\ \frac{0.1380}{4.601}\right] = [0.00421;\ 0.02998].$$

To obtain the confidence interval for the standard deviation, σ, we take the square root:

$$[\sqrt{0.00421};\ \sqrt{0.02998}] = [0.065;\ 0.173].$$

Exam-style questions

c) A sample of $n = 25$ bolts gives the following diameters (in mm):

12.57	12.28	12.67	12.71	12.09
12.23	12.54	12.44	12.50	12.32
12.70	12.68	12.52	12.75	12.61
12.32	12.59	12.30	12.70	12.50
12.47	12.36	12.77	12.47	12.41

From the data: $\bar{x} = 12.50$ mm and $s = 0.18$ mm, assumed normally distributed.
The following computations were performed in Python:

```
print(12.50 - stats.t.ppf(0.975, df=24) * 0.18/np.sqrt
    (25))
12.43

print(12.50 + stats.t.ppf(0.975, df=24) * 0.18/np.sqrt
    (25))
12.57
```

Which of the following is the correct interpretation of these results?

1) 97.5% of all possible sample means from repeated sampling will have values in the interval $[12.43, 12.57]$.
2) 97.5% of individual bolts have diameters in the interval $[12.43, 12.57]$.
3) There is a 95% probability that a newly manufactured bolt's diameter falls in the interval $[12.43, 12.57]$.
4) The true mean μ falls in the interval $[12.43, 12.57]$ with certainty.
5) The interval $[12.43, 12.57]$, is the 95% confidence interval for the population mean μ.

|||| Solution

Correct answer: **5)**.

3.3 Concrete items (hypothesis testing)

Exercise 3.3 Concrete items (hypothesis testing)

This is a continuation of Exercise 1, so the same setting and data is used (read the initial text of it).

a) To investigate whether the requirement to the mean is fulfilled (with $\alpha = 5\%$), test the following hypothesis:

$$H_0 : \mu = 3000$$
$$H_1 : \mu \neq 3000.$$

What is the evidence against the null hypothesis?
(Quantile: $t_{0.975,8} = 2.306$)

Solution

The observed t-statistic is:

$$t_{\text{obs}} = \frac{3002.333 - 3000}{3.708/\sqrt{9}} = 1.885.$$

The p-value (using t-distribution with $\nu = 8$ degrees of freedom):

$$p\text{-value} = 2 \cdot P(T > 1.885) = 0.096.$$

There is weak evidence against the null hypothesis. At $\alpha = 0.05$ (critical value ± 2.306) we do not reject H_0; the mean length is consistent with 3000 mm.

```
t_obs = (3002.333 - 3000) / (3.708 / np.sqrt(9))
print(t_obs, 2*(1 - stats.t.cdf(t_obs, df=8)))
```

b) What would the level $\alpha = 0.01$ critical values be for this test, and how are they interpreted?
(Quantile: $t_{0.995,8} = 3.355$)

|||| Solution

The critical values are $\pm t_{0.995,8} = \pm 3.355$. For a new experiment, the observed t-statistic must exceed 3.355 in absolute value to lead to a significant result at $\alpha = 0.01$.

c) What would the level $\alpha = 0.05$ critical values be for this test? Compare with the values found in the previous question.
(Quantile: $t_{0.975,8} = 2.306$)

|||| Solution

The critical values are $\pm t_{0.975,8} = \pm 2.306$. Compared to the $\alpha = 0.01$ case, it is easier to detect an effect at significance level $\alpha = 0.05$ since the critical value is smaller.

d) Investigate, by means of a normal QQ plot, whether the data appears to come from a normal distribution.

|||| Solution

The nine data points do not deviate from the reference line more than what is typical for truly normally distributed samples of size $n = 9$. We cannot falsify the normality assumption; the data are consistent with a normal distribution.

e) What exactly is plotted in the normal QQ plot? Specifically, what are the x- and y-coordinates of the points?

|||| Solution

The y-coordinates are the ordered observations $x_{(1)} \leq \cdots \leq x_{(9)}$: 2997, 2998, 2999, 3001, 3003, 3005, 3005, 3006, 3007.

The x-coordinates are theoretical standard normal quantiles. For $n \leq 10$ the formula is $p_i = (i - 3/8)/(n + 1/4)$, giving $z = \Phi^{-1}(p_i)$ for $i = 1, \ldots, 9$.

Exam-style questions

f) A tile manufacturer tests $H_0 : \mu = 200$ vs. $H_1 : \mu \neq 200$ (tile thickness in mm) using $n = 14$ tiles, finding $t_{\text{obs}} = 2.05$. The critical value at $\alpha = 0.05$ is $t_{0.975,13} = 2.160$, and at $\alpha = 0.10$ it is $t_{0.95,13} = 1.771$. What is the correct conclusion?

1) H_0 is not rejected at $\alpha = 0.05$ (since $|t_{\text{obs}}| = 2.05 < 2.160$), but *is* rejected at $\alpha = 0.10$ (since $2.05 > 1.771$).
2) H_0 is rejected at both $\alpha = 0.05$ and $\alpha = 0.10$, since t_{obs} is positive.
3) H_0 is not rejected at either significance level, since $t_{\text{obs}} < 3$.
4) H_0 is rejected at $\alpha = 0.05$ but not at $\alpha = 0.10$, since a smaller α means a smaller (easier to exceed) critical value.
5) The conclusion cannot be determined without also knowing the exact p-value.

|||| Solution

Correct answer: **1)**. Compare $|t_{\text{obs}}| = 2.05$ to each critical value: it is smaller than 2.160 (so H_0 stands at $\alpha = 0.05$) but larger than 1.771 (so H_0 is rejected at $\alpha = 0.10$) — exactly the comparison logic used in the previous two questions of this exercise, just requiring the conclusion to be drawn rather than stated. Option 4) has the direction backwards: a *smaller* α requires *stronger* evidence, hence a *larger* critical value, not a smaller one. Option 5) is unnecessary — comparing t_{obs} directly to the critical values is sufficient to reach a conclusion.

g) For a normal QQ-plot of $n = 12$ ordered observations, using the same plotting-position formula as above ($p_i = (i - 3/8)/(n + 1/4)$), what is the theoretical (horizontal-axis) coordinate for the *smallest* observation ($i = 1$)?

1) $p_1 = 0.0510$, giving $z_1 = \Phi^{-1}(0.0510) \approx -1.64$
2) $p_1 = 1/12 = 0.0833$, giving $z_1 \approx -1.38$
3) $p_1 = (1 - 3/8)/12 = 0.0521$, giving $z_1 \approx -1.63$
4) $z_1 = 0.0510$
5) $p_1 = (12 - 3/8)/(12 + 1/4) = 0.949$, giving $z_1 \approx 1.64$

|||| Solution

Correct answer: **1)**. $p_1 = (1 - 3/8)/(12 + 1/4) = 0.625/12.25 = 0.0510$, and $z_1 = \Phi^{-1}(0.0510) \approx -1.64$. Option 2) uses the naive (uncorrected) formula i/n instead of the plotting-position formula. Option 3) forgets to add $1/4$ to the denominator. Option 4) confuses the probability p_1 itself for the z-quantile. Option 5) mistakenly uses $i = n$ (the *largest* observation's position) instead of $i = 1$.

```
print(stats.norm.ppf((1 - 3/8) / (12 + 1/4)))
```

3.4 Aluminum profile (hypothesis testing)

|||| Exercise 3.4 Aluminum profile (hypothesis testing)

We use the same setting and data as in Exercise 2:

The length of an aluminum profile is checked by taking a sample of 16 items whose lengths are measured. The measurement results from this sample are listed below, all measurements are in mm:

180.02	180.00	180.01	179.97	179.92	180.05	179.94	180.10
180.24	180.12	180.13	180.22	179.96	180.10	179.96	180.06

From the data: $\bar{x} = 180.05$ mm and $s = 0.0959$ mm.

It can be assumed that the sample comes from a normally distributed population.

a) Find the evidence against the following hypothesis:

$$H_0 : \mu = 180.$$

Use Python to compute quantiles and probabilities from relevant probability distribution(s).

|||| Solution

We use Method 3.23.

First we compute the observed t-statistic:

$$t_{\text{obs}} = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} = \frac{180.05 - 180}{0.0959/\sqrt{16}} = 2.085.$$

t_{obs} represents our *observation* of the stochastic variable T, which is expected to follow a t-distribution with $\nu = n - 1$ degrees of freedom (see also Theorem 3.5).

We now calculate the p-value: the probability of obtaining a value of T that is at least as extreme (at least this far away from zero) as the test statistic that was actually observed.

The p-value (using t-distribution with $\nu = 15$ degrees of freedom):

$$p\text{-value} = P(T < -2.085) + P(T > 2.085) = 2 \cdot P(T < -2.085)$$

The value is computed using Python:

```
print(2 * stats.t.ppf(-2.085, df = 15)
0.0546
```

So we have p-value $= 0.0546$.

There is weak evidence against H_0 (see Table 3.1 in the book).

The test statistic and p-value could also be calculated in Python (after having saved the raw data in the variable x) in the following way:

```
t_obs, pval = stats.ttest_1samp(x, popmean=180)
print(t_obs, pval)
2.085 0.0546
```

b) If the hypothesis test

$$H_0 : \mu = 180,$$
$$H_1 : \mu \neq 180,$$

is carried out, what are the level $\alpha = 1\%$ critical values?

Use Python to compute quantiles and probabilities from relevant probability distribution(s).

|||| Solution

The critical values (see Definition 3.31) are the 0.005- and 0.995-quantiles of the t-distribution with $\nu = 15$ degrees of freedom.

We use Python to compute the values:

```
print(stats.t.ppf(0.005, df=15))
-2.94671

print(stats.t.ppf(0.995, df=15))
2.94671
```

This means that if the observed t_{obs} falls within the interval $[-2.94671; 2.94671]$, then we fail to reject the null hypothesis as the 1% significance level (which is the case here since $t_{obs} = 2.085$). This is not surprising since we have already seen that the p-value is indeed larger than 0.01.

c) What is the 99% confidence interval for μ?

Use Python to compute quantiles and probabilities from relevant probability distribution(s).

|||| Solution

To compute the 99% confidence interval we need the quantile $t_{0.995,15} = 2.947$, already found in question b. We use Method 3.9:

$$180.05 \pm 2.947 \cdot \frac{0.0959}{\sqrt{16}} = [179.98;\ 180.12].$$

d) Carry out the hypothesis test

$$H_0 : \mu = 180,$$
$$H_1 : \mu \neq 180,$$

using $\alpha = 5\%$.

|||| Solution

Since we have already found the p-value $= 0.0546$ and concluded that $0.0546 > 0.05 = \alpha$, we fail to reject the null hypothesis at the 5% significance level. In other words, the data do not provide sufficient evidence to conclude that μ is significantly different from 180 mm.

Exam-style questions

e) A textile factory wants to test the thread tensile strength (measured in Newtons) by performing the hypothesis test:

$$H_0 : \mu = 500 \quad \text{vs} \quad H_1 : \mu \neq 500$$

at significance level $\alpha = 5\%$.

The thread tensile strength is then measured in $n = 20$ textile samples.

Without performing the full test, a 95% confidence interval for μ is computed as $[497.2, 503.8]$.

What can be concluded about the hypothesis test?

1) The confidence interval only tells us about μ in general, so we cannot conclude anything about the specific null hypothesis.
2) Since 500 lies inside the interval, the null hypothesis H_0 is rejected at significance level $\alpha = 2.5\%$. We cannot conclude anything about the conclusion at significance level $\alpha = 5\%$.
3) Since 500 lies inside the interval, the null hypothesis H_0 is rejected at significance level $\alpha = 5\%$.
4) Since 500 lies inside the interval, the null hypothesis H_0 cannot be rejected at $\alpha = 5\%$.
5) We would need to know $\bar{x}$ and s separately to determine whether H_0 is rejected.

|||| Solution

Correct answer: **4)**. A two-sided hypothesis test at level α and a $(1-\alpha)$ confidence interval are two views of the same calculation: $H_0 : \mu = \mu_0$ is rejected at level α if and only if μ_0 falls *outside* the $(1-\alpha)$ confidence interval for μ. Since $500 \in [497.2, 503.8]$, H_0 is not rejected.

3.5 Transport times

|||| Exercise 3.5 Transport times

A company, MM, selling items online wants to compare the transport times for two transport firms for delivery of the goods. To compare the two companies, recordings of delivery times on a specific route were made, with a sample size of $n = 9$ for each firm. The following data (in days) were found:

$$\text{Firm A: } 2.20,\ 1.58,\ 2.41,\ 2.49,\ 1.15,\ 1.46,\ 2.12,\ 1.91,\ 2.05,$$
$$\text{Firm B: } 0.97,\ 1.07,\ 1.02,\ 1.34,\ 2.33,\ 0.88,\ 2.09,\ 2.24,\ 1.47.$$

This gives $\bar{y}_A = 1.93$ d and $s_A = 0.45$ d, and $\bar{y}_B = 1.49$ d and $s_B = 0.58$ d. Note that d is the SI unit for days. It is assumed that data can be regarded as stemming from normal distributions.

a) We want to test the hypothesis

$$H_0 : \mu_A = \mu_B$$
$$H_1 : \mu_A \neq \mu_B.$$

What is the p-value, interpretation and conclusion for this test (at $\alpha = 5\%$)?

|||| Solution

This is the Welch two-sample t-test. The test statistic is:

$$t_{\text{obs}} = \frac{1.93 - 1.49}{\sqrt{0.45^2/9 + 0.58^2/9}} = 1.8.$$

The degrees of freedom:

$$\nu = \frac{\left(\frac{0.45^2}{9} + \frac{0.58^2}{9}\right)^2}{\frac{(0.45^2/9)^2}{8} + \frac{(0.58^2/9)^2}{8}} = 15.0.$$

The p-value using a t-distribution with $\nu = 15.0$:

$$p\text{-value} = 2 \cdot P(T > 1.8) = 0.0922.$$

There is weak evidence against the null hypothesis. At $\alpha = 0.05$ we cannot reject H_0; the two firms are not shown to differ significantly.

```
yA = np.array([2.20, 1.58, 2.41, 2.49, 1.15, 1.46, 2.12,
    1.91, 2.05])
yB = np.array([0.97, 1.07, 1.02, 1.34, 2.33, 0.88, 2.09,
    2.24, 1.47])

t_obs, pval = stats.ttest_ind(yA, yB, equal_var=False)
print(t_obs, pval)

1.798 0.09219
```

b) Find the 95% confidence interval for the mean difference $\mu_A - \mu_B$.
(One of the following t-quantiles may be useful: $t_{0.95,15} = 1.753$, $t_{0.975,15} = 2.131$, $t_{0.975,16} = 2.120$, $t_{0.995,15} = 2.947$*)*

|||| Solution

The correct quantile is $t_{0.975,15} = 2.131$.

$$(1.93 - 1.49) \pm 2.131\sqrt{0.45^2/9 + 0.58^2/9} = [-0.082;\ 0.962].$$

c) What is the power of a study with $n = 9$ observations in each of the two samples for detecting a potential mean difference of 0.4 between the firms (assume $\sigma = 0.5$ and $\alpha = 0.05$)?

|||| Solution

```
delta = smp.TTestIndPower().solve_power(nobs1=9, power=0.8,
    alpha=0.05) * 0.5
print(delta)

0.703
```

A mean difference of approximately 0.70 d is detectable with probability 0.8.

d) What effect size (mean difference) could be detected with $n = 9$ observations in each sample with a power of 0.8 (assume $\sigma = 0.5$ and $\alpha = 0.05$)?

|||| Solution

```
delta = TTestIndPower().solve_power(nobs1=9, power=0.8,
    alpha=0.05)*0.5
print(delta)
0.703
```

A mean difference of approximately 0.70 d is detectable with probability 0.8.

e) How large a sample size from each firm is needed to detect a potential mean difference of 0.4 with probability 0.90 (assume $\sigma = 0.5$ and $\alpha = 0.05$)?

|||| Solution

```
n = smp.TTestIndPower().solve_power(effect_size=0.4/0.5,
    power=0.90, alpha=0.05)
print(np.ceil(n))

34.0
```

$n = 34$ observations from each firm are needed.

Exam-style questions

f) Holding α, σ, and the true effect size fixed, which of the following statements about statistical power is false?

1) Increasing the sample size n increases power.
2) Increasing α increases power (at the cost of a higher Type I error rate).
3) A larger true effect size (mean difference) gives higher power for the same n.

4) Power and the probability of a Type II error (β) always sum to exactly 1, so increasing power always decreases β correspondingly.

5) Power does not depend on the sample size n, only on α and the effect size.

|||| Solution

Correct answer: **5)**. Statement 5) is false — indeed the opposite is the whole point of the last two questions of this exercise: power increases with n (all else equal), which is exactly why a larger sample is needed to reliably detect a smaller effect size. Statements 1)-4) are all true general facts about power: it increases with n, with α, and with the true effect size, and by definition power$= 1 - \beta$.

3.6 Cholesterol

|||| Exercise 3.6 Cholesterol

In a clinical trial of a cholesterol-lowering agent, 15 patients' cholesterol (in mmol/L) has been measured before treatment and 3 weeks after starting treatment. Data are listed in the following table:

Patient	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Before	9.1	8.0	7.7	10.0	9.6	7.9	9.0	7.1	8.3	9.6	8.2	9.2	7.3	8.5	9.5
After	8.2	6.4	6.6	8.5	8.0	5.8	7.8	7.2	6.7	9.8	7.1	7.7	6.0	6.6	8.4

a) Can there, based on these data, be demonstrated a significant decrease in cholesterol levels with $\alpha = 0.001$?

|||| Solution

This is a *paired* setting (same patients measured twice). The differences $D_i = \text{Before}_i - \text{After}_i$ are used. A paired t-test gives a (non-directional) p-value of 0.0000037, providing very strong evidence against the null hypothesis. We conclude beyond any reasonable doubt that the mean cholesterol level has decreased after 3 weeks.

```
x1 = np.array
   ([9.1,8.0,7.7,10.0,9.6,7.9,9.0,7.1,8.3,9.6,8.2,9.2,7.3,8.5,9.5])

x2 = np.array
   ([8.2,6.4,6.6,8.5,8.0,5.8,7.8,7.2,6.7,9.8,7.1,7.7,6.0,6.6,8.4])


result = stats.ttest_rel(x1, x2)
t_obs, p_val, df = result.statistic, result.pvalue, result.
   df
print(t_obs, p_val, df)

7.3406 3.672e-06 14
```

Exam-style questions

b) A running coach measures 10 athletes' 5K times before and after a 12-week training program. Which test is appropriate to determine whether the program significantly changed average times, and why?

1) A paired t-test on the differences $D_i = \text{Before}_i - \text{After}_i$, since the same athletes are measured twice and their before/after times are not independent of each other.
2) An independent two-sample t-test, since "before" and "after" are two separate columns of measurements.
3) A one-sample t-test on the "after" times alone, ignoring the "before" measurements entirely.
4) A paired t-test on the ratios $\text{Before}_i/\text{After}_i$ instead of the differences.
5) It does not matter which test is used; paired and independent two-sample tests always give the same p-value.

|||| Solution

Correct answer: **1)**. Exactly as in the cholesterol study above, the same individuals are measured twice, so the before/after measurements are dependent (an athlete who is fast before training tends to be fast after training too). A paired test on the differences removes this between-athlete variability, leaving only the effect of training — which is why the paired test is more powerful here than an (incorrectly applied) independent two-sample test, which is what option 2) and option 5) wrongly suggest could be used interchangeably.

3.7 Pulse

|||| Exercise 3.7 Pulse

13 runners had their pulse measured at the end of a workout and again 1 minute later, with the following results:

Runner	1	2	3	4	5	6	7	8	9	10	11	12	13
Pulse end	173	175	174	183	181	180	170	182	188	178	181	183	185
Pulse 1 min	120	115	122	123	125	140	108	133	134	121	130	126	128

From the data: $\bar{x}_{\text{end}} = 179.46$, $\bar{x}_{1\text{min}} = 125.00$, $s_{\text{end}} = 5.190$, $s_{1\text{min}} = 8.956$, $s_D = 5.768$.

a) What is the 99% confidence interval for the mean pulse drop (the drop during 1 minute from end of workout)?
(One of the following t-quantiles may be useful: $t_{0.975,12} = 2.179$, $t_{0.995,12} = 3.054$, $t_{0.995,13} = 3.012$*)*

|||| Solution

This is a paired setting: each runner is measured twice, so we work with the differences $D_i = \text{Pulse end}_i - \text{Pulse 1 min}_i$ (the pulse drop during that minute). With $n = 13$, $\bar{x}_D = 54.46$, $s_D = 5.768$ and $t_{0.995,12} = 3.054$, a 99% CI for the mean pulse drop μ_D is

$$\bar{x}_D \pm t_{0.995,12} \cdot \frac{s_D}{\sqrt{n}} = 54.46 \pm 3.054 \cdot \frac{5.768}{\sqrt{13}} = [49.58;\ 59.35].$$

```
x1 = np.array
    ([173,175,174,183,181,180,170,182,188,178,181,183,185])
x2 = np.array
    ([120,115,122,123,125,140,108,133,134,121,130,126,128])

CI = stats.ttest_rel(x1, x2).confidence_interval(0.99)
print(CI)

ConfidenceInterval(low=49.575065928685966, high
    =59.348010994390954)
```

b) Consider the 13 pulse end measurements (first row in the table). What is the 95% confidence interval for the standard deviation of these?
(Some of the following χ^2*-quantiles may be useful:* $\chi^2_{0.975,\,12} = 23.34$, $\chi^2_{0.025,\,12} = 4.40$, $\chi^2_{0.975,\,13} = 24.74$, $\chi^2_{0.025,\,13} = 5.01$*)*

|||| Solution

The 95% CI for σ is based on

$$\left[\sqrt{\frac{(n-1)s^2_{\text{end}}}{\chi^2_{0.975,\,n-1}}};\ \sqrt{\frac{(n-1)s^2_{\text{end}}}{\chi^2_{0.025,\,n-1}}}\right].$$

With $n = 13$, $s_{\text{end}} = 5.190$, $(n-1)s^2_{\text{end}} = 12 \times 5.190^2 = 323.2$, $\chi^2_{0.975,12} = 23.34$ and $\chi^2_{0.025,12} = 4.40$:

$$\left[\sqrt{\frac{323.2}{23.34}};\ \sqrt{\frac{323.2}{4.40}}\right] = [3.72;\ 8.57].$$

We are 95% confident that $\sigma \in [3.72;\ 8.57]$.

Exam-style questions

c) In the pulse exercise above, note that $s_D = 5.768$ is *not* equal to $s_{\text{end}} - s_{1\text{min}} = 5.190 - 8.956 = -3.766$, nor to $\sqrt{s^2_{\text{end}} + s^2_{1\text{min}}} = \sqrt{5.190^2 + 8.956^2} = 10.35$. Why?

1) Because the end-of-workout and 1-minute pulse measurements for the same runner are correlated (a runner with a high end-pulse tends to have a high 1-min pulse too), so s_D must be computed directly from the individual differences D_i, not derived from s_{end} and $s_{1\text{min}}$ alone via a simple formula.

2) s_D is simply a typo in the exercise; it should equal $s_{\text{end}} - s_{1\text{min}}$.

3) s_D should always equal $\sqrt{s^2_{\text{end}} + s^2_{1\text{min}}}$, and the discrepancy is a rounding error.

4) s_D can never be derived from s_{end} and $s_{1\text{min}}$ under any circumstances, even if the correlation between them were also known.

5) s_D only differs because the two rows have different sample sizes.

|||| Solution

Correct answer: **1)**. For paired (dependent) measurements, $\mathrm{V}(D) = \mathrm{V}(\text{end}) + \mathrm{V}(1\text{min}) - 2\,\mathrm{Cov}(\text{end}, 1\text{min})$ — the variance-addition rule from independent variables picks up an extra covariance term once the two measurements are correlated, which is exactly the case here (same runner, measured twice). This is why s_D must be computed from the differences directly. Option 4) is close but overstates the case: if the correlation r were also known, s_D *could* in principle be recovered via $s_D = \sqrt{s_{\text{end}}^2 + s_{1\text{min}}^2 - 2rs_{\text{end}}s_{1\text{min}}}$ — it is only the naive formulas in options 2) and 3) that are wrong, not the general idea. Option 5) is false; both rows have $n = 13$.

3.8 Foil production

Exercise 3.8 Foil production

In the production of a certain foil (film), the foil is controlled by measuring the thickness in a number of points distributed over the width. The production is considered stable if the mean of the difference between the maximum and minimum measurements does not exceed 0.35 mm. At a given day, the following random samples are observed for 10 foils:

Foil	1	2	3	4	5	6	7	8	9	10
Max. (mm)	2.62	2.71	2.18	2.25	2.72	2.34	2.63	1.86	2.84	2.93
Min. (mm)	2.14	2.39	1.86	1.92	2.33	2.00	2.25	1.50	2.27	2.37
Max−Min (D)	0.48	0.32	0.32	0.33	0.39	0.34	0.38	0.36	0.57	0.56

The following statistics may potentially be used:

$$\bar{y}_{\text{max}} = 2.508,\ \bar{y}_{\text{min}} = 2.103,\ s_{y_{\text{max}}} = 0.3373,\ s_{y_{\text{min}}} = 0.2834,\ s_D = 0.09664.$$

a) What is a 95% confidence interval for the mean difference μ_D?
(One of the following t-quantiles may be useful: $t_{0.95,9} = 1.833$, $t_{0.975,9} = 2.262$, $t_{0.975,10} = 2.228$, $t_{0.995,9} = 3.250$)

Solution

The mean difference is $\bar{y}_D = \bar{y}_{\text{max}} - \bar{y}_{\text{min}} = 0.405$, with $s_D = 0.09664$ and $n = 10$. Using $t_{0.975,9} = 2.262$:

$$\bar{y}_D \pm t_{0.975,9} \cdot \frac{s_D}{\sqrt{n}} = 0.405 \pm 2.262 \cdot \frac{0.09664}{\sqrt{10}} = [0.336;\ 0.474].$$

Notice that the CI contains $\mu_D = 0.35$, so we already know the next question's hypothesis will not be rejected.

```
x1 = np.array([2.62, 2.71, 2.18, 2.25, 2.72, 2.34, 2.63,
    1.86, 2.84, 2.93])
x2 = np.array([2.14, 2.39, 1.86, 1.92, 2.33, 2.00, 2.25,
    1.50, 2.27, 2.37])
```

```
CI = stats.ttest_rel(x1, x2).confidence_interval(0.95)
print(CI)

ConfidenceInterval(low=0.3358, high=0.4741)
```

b) How much evidence is there that the mean difference is different from 0.35? State the null hypothesis, t-statistic and p-value for this question.

|||| Solution

We test $H_0 : \mu_D = 0.35$ against $H_1 : \mu_D \neq 0.35$. The test statistic is

$$t_{\text{obs}} = \frac{\bar{y}_D - 0.35}{s_D/\sqrt{n}} = \frac{0.405 - 0.35}{0.09664/\sqrt{10}} = 1.80,$$

which, compared to a t-distribution with $n - 1 = 9$ degrees of freedom, gives

$$p\text{-value} = 2 \cdot P(T > 1.80) = 0.105.$$

Since $p = 0.105 > 0.05$, there is little or no evidence against H_0: we cannot conclude that the mean difference deviates from 0.35. This agrees with the previous question, where the 95% CI, [0.336; 0.474], contained 0.35.

```
diff_data = x1 - x2

t_obs, p_val = stats.ttest_1samp(diff_data, popmean=0.35)
print(t_obs, p_val)

1.799 0.105
```

Exam-style questions

c) A quality engineer wants the 95% CI for a mean difference (like μ_D above) to have a margin of error no larger than 0.03 mm. From a pilot study, $s_D \approx 0.095$ mm. Using the (approximate, large-sample) formula $n \approx \left(\frac{z_{0.975}\, s_D}{ME}\right)^2$ with $z_{0.975} = 1.960$, what is the minimum sample size needed?

1) $n = 39$
2) $n = 7$
3) $n = 10$
4) $n = 28$
5) $n = 38$

|||| Solution

Correct answer: **1)**.

$$n \approx \left(\frac{1.960 \times 0.095}{0.03}\right)^2 = (6.207)^2 = 38.5,$$

which must be rounded *up* to $n = 39$ to guarantee the margin of error requirement is met. Option 2) forgets to square the ratio. Option 3) mistakenly uses the full interval width ($2 \times ME = 0.06$) instead of the margin of error itself. Option 4) uses $z_{0.95} = 1.645$ instead of the correct $z_{0.975} = 1.960$ for a 95% interval. Option 5) makes the right calculation but rounds *down* instead of up, which would not quite guarantee the desired precision.

3.9 Course project

Exercise 3.9 Course project

At a specific education it was decided to introduce a project, running through the course period, as part of the grade point evaluation. In order to assess whether it has changed the percentage of students passing the course, the following data was collected:

	Before introduction of project	After introduction of project
Number of students evaluated	50	24
Number of students failed	13	3
Average grade point $\bar{x}$	6.420	7.375
Sample standard deviation s	2.205	1.813

a) Assuming that the grades are approximately normally distributed in each group, test the hypothesis:

$$H_0 : \mu_{\text{Before}} = \mu_{\text{After}},$$
$$H_1 : \mu_{\text{Before}} \neq \mu_{\text{After}}.$$

State the test statistic, p-value and conclusion.

Solution

Since the two groups have quite different sample sizes and standard deviations, we use Welch's t-test:

$$t_{\text{obs}} = \frac{\bar{x}_{\text{Before}} - \bar{x}_{\text{After}}}{\sqrt{s^2_{\text{Before}}/n_{\text{Before}} + s^2_{\text{After}}/n_{\text{After}}}} = \frac{6.420 - 7.375}{\sqrt{2.205^2/50 + 1.813^2/24}} = -1.97.$$

The degrees of freedom are also found from the following equation:

$$\nu = \frac{\left(\dfrac{s^2_{\text{Before}}}{n_{\text{Before}}} + \dfrac{s^2_{\text{After}}}{n_{\text{After}}}\right)^2}{\dfrac{(s^2_{\text{Before}}/n_{\text{Before}})^2}{n_{\text{Before}} - 1} + \dfrac{(s^2_{\text{After}}/n_{\text{After}})^2}{n_{\text{After}} - 1}} = \frac{(0.09724 + 0.13696)^2}{0.09724^2/49 + 0.13696^2/23} = \frac{0.05485}{0.001009} \approx 54.4.$$

This gives a non-directional p-value:

$$p\text{-value} = 2 \cdot P(T > 1.97) = 0.054.$$

At $\alpha = 5\%$ we cannot reject H_0: there is no significant difference in mean grade point before and after the introduction of the project.

b) A 99% confidence interval for the mean grade point difference becomes?
(One of the following t-quantiles may be useful: $t_{0.975,54.4} = 2.005$, $t_{0.995,54.4} = 2.669$, $t_{0.995,72} = 2.646$*)*

|||| Solution

The correct quantile is $t_{0.995,54.4} = 2.669$. The 99% CI for $\mu_{\text{Before}} - \mu_{\text{After}}$ is

$$(\bar{x}_{\text{Before}} - \bar{x}_{\text{After}}) \pm t_{0.995,54.4}\sqrt{\frac{s^2_{\text{Before}}}{n_{\text{Before}}} + \frac{s^2_{\text{After}}}{n_{\text{After}}}}.$$

With $\bar{x}_{\text{Before}} - \bar{x}_{\text{After}} = 6.420 - 7.375 = -0.955$ and $t_{0.995,54.4} = 2.669$:

$$-0.955 \pm 2.669\sqrt{2.205^2/50 + 1.813^2/24} = [-2.247;\ 0.337].$$

We are 99% confident that the true mean difference $\mu_{\text{Before}} - \mu_{\text{After}}$ lies in $[-2.247;\ 0.337]$. Note that, as expected, the interval contains 0, consistent with the non-rejection of H_0 in the previous question.

c) A 95% confidence interval for the grade point standard deviation *after* the introduction of the project becomes?
Some of the following χ^2*-quantiles may be useful:* $\chi^2_{0.975,23} = 38.076$, $\chi^2_{0.025,23} = 11.689$, $\chi^2_{0.975,24} = 39.364$, $\chi^2_{0.025,24} = 12.401$, $\chi^2_{0.95,23} = 35.172$, $\chi^2_{0.05,23} = 13.091$.

|||| Solution

The correct quantiles are $\chi^2_{0.975,23} = 38.076$ and $\chi^2_{0.025,23} = 11.689$ Using $(n-1)s^2 = 23 \times 1.813^2 = 23 \times 3.287 = 75.6$:

$$\left[\sqrt{\frac{(n-1)s^2_{\text{After}}}{\chi^2_{0.975,n-1}}};\ \sqrt{\frac{(n-1)s^2_{\text{After}}}{\chi^2_{0.025,n-1}}}\right] = \left[\frac{\sqrt{23 \times 1.813^2}}{\sqrt{38.076}};\ \frac{\sqrt{23 \times 1.813^2}}{\sqrt{11.689}}\right] = [1.41;\ 2.54].$$

We accept that $\sigma_{\text{After}} \in [1.41;\ 2.54]$.

Exam-style questions

d) The table above also reports failure counts: 13 of 50 students failed before the project was introduced, and 3 of 24 failed after. If the education wants to test whether the *failure proportion* (not the mean grade) has changed, which of the following is the appropriate approach?

1) A two-sample test for proportions (comparing $\hat{p}_{\text{before}} = 13/50 = 0.26$ and $\hat{p}_{\text{after}} = 3/24 = 0.125$ using a z-test), not the Welch t-test used for the mean grades.
2) The same Welch t-test used for the mean grades, applied directly to the counts 13 and 3.
3) A paired t-test, since the counts come from the same course.
4) No test is needed; since $0.26 \neq 0.125$, the proportions are automatically significantly different.
5) A chi-squared goodness-of-fit test comparing the counts to a uniform distribution.

|||| Solution

Correct answer: **1)**. The mean-grade columns are continuous data, appropriate for a (Welch) t-test — but the failure counts are binary (pass/fail) data, which call for a proportions test instead. Choosing the right test starts with recognising the data type, not just plugging numbers into a familiar formula. Option 2) misapplies the t-test formula to count data it wasn't designed for. Option 4) ignores sampling variability entirely — an observed difference is never automatically significant regardless of sample size.

3.10 Concrete items (sample size)

|||| Exercise 3.10 Concrete items (sample size)

This is a continuation of Exercise 1, so the same setting and data is used (read the initial text of it).

a) A study is planned for a new supplier. It is expected that the standard deviation will be approximately $\sigma = 3$ mm. We want a 90% confidence interval for the mean value in this new study to have a width of 2 mm. How many items should be sampled to achieve this?
(Quantile: $z_{0.95} = 1.645$*)*

|||| Solution

Using the sample size formula with wanted margin of error $ME = 1$ (half the width):

$$n = \left(\frac{z_{0.95} \cdot \sigma}{ME}\right)^2 = \left(\frac{1.645 \times 3}{1}\right)^2 = 24.35.$$

Hence at least $n = 25$ items must be sampled.

b) Answer the sample size question above but requiring the 99% confidence interval to have the same width of 2 mm.
(Quantile: $z_{0.995} = 2.576$*)*

|||| Solution

$$n = \left(\frac{z_{0.995} \times 3}{1}\right)^2 = \left(\frac{2.576 \times 3}{1}\right)^2 = 59.72.$$

Hence at least $n = 60$ items must be sampled.

c) (Challenging) For the two sample sizes found above, find the probability that the corresponding confidence interval in a future study will be more than 10% wider than planned (still assuming $\sigma^2 = 9$).

|||| Solution

The random half-width of the CI is $t_{1-\alpha/2} \cdot S/\sqrt{n}$. Since $(n-1)S^2/9 \sim \chi^2_{n-1}$, the probability that the half-width exceeds $1.1 \times ME$ can be written as a tail probability of the χ^2 distribution.

For $\alpha = 0.10$ and $n = 25$ (so $t_{0.95,24} = 1.711$):

$$P\left(t_{0.95} \cdot \frac{S}{\sqrt{25}} > 1.1\right) = P\left(\chi^2_{24} > \frac{1.1^2 \times 24 \times 25}{1.711^2 \times 9}\right) = P(\chi^2_{24} > 27.56) = 0.28.$$

Almost 30% of the time the planned experiment will yield a CI wider than intended.

For $\alpha = 0.01$ and $n = 60$ (so $t_{0.995,59} = 2.662$):

$$P(\chi^2_{59} > 68.0) \approx 0.22.$$

In this case it only happens in about 22% of cases.

d) With the sample size $n = 25$ from the first question, what are the probabilities of detecting (i.e. rejecting $H_0 : \mu = 3000$) when the true mean is $\mu_1 = 3001, 3002$, or 3003, respectively? Assume $\alpha = 0.05$ and $\sigma = 3$.

|||| Solution

Using the power of a one-sample t-test:

$$\delta = |\mu_1 - \mu_0| \in \{1, 2, 3\}.$$

```
from statsmodels.stats.power import TTestPower
for d in [1, 2, 3]:
    print(d, TTestPower().power(d/3, nobs=25, alpha=0.05))
```

The three powers are approximately 0.36, 0.89 and 0.998. A difference of 1 mm is unlikely to be detected, but differences of 2 and 3 mm have high detection probability.

e) Answer the same question as above but using $n = 60$.

|||| Solution

```
for d in [1, 2, 3]:
    print(d, TTestPower().power(d/3, nobs=60, alpha=0.05))
```

The three powers are approximately 0.72, 0.999 and 1.00. With $n = 60$ even a difference of 1 mm has a reasonable detection probability.

f) What sample size is needed to achieve a power of 0.80 for detecting an effect of size 0.5 mm (assume $\sigma = 3$ and $\alpha = 0.05$)?

|||| Solution

```
n = TTestPower().solve_power(effect_size=0.5/3, power=0.8,
    alpha=0.05)
print(np.ceil(n))
```

Using the formula: $n = \left(3 \cdot \frac{z_{0.8}+z_{0.975}}{0.5}\right)^2 \approx 283$, or more precisely $n = 285$ from the t-test power function. A difference of 0.5 mm is expensive to detect reliably.

g) Assume you only have finances for $n = 50$ observations. How large a difference would you be able to detect with probability 0.80?

|||| Solution

```
delta = TTestPower().solve_power(nobs=50, power=0.8, alpha
    =0.05)*3
print(delta)
```

The detectable effect is approximately 1.21 mm. A true mean of 3001.2 or higher (or 2998.8 or lower) would be detected with probability 0.80.

Exam-style questions

h) A new supplier's pilot study suggests $\sigma \approx 5$ mm for a different concrete item. The company wants a 95% CI for the mean to have a total width of 4 mm. Using $z_{0.975} = 1.960$, how many items should be sampled?

1) $n = 25$
2) $n = 7$
3) $n = 17$
4) $n = 5$
5) $n = 601$

|||| Solution

Correct answer: **1)**. The margin of error is half the total width: $ME = 4/2 = 2$ mm. Then

$$n = \left(\frac{z_{0.975}\,\sigma}{ME}\right)^2 = \left(\frac{1.960 \times 5}{2}\right)^2 = 24.0,$$

rounded up to $n = 25$. Option 2) mistakenly uses the full width (4) as if it were the margin of error, instead of half of it. Option 3) uses $z_{0.95} = 1.645$ instead of the correct $z_{0.975} = 1.960$ for a 95% interval. Option 4) forgets to square the ratio. Option 5) mistakenly plugs in $\sigma^2 = 25$ where $\sigma = 5$ belongs.

i) The exercise above found that for detecting $\mu_1 = 3001$ (a 1 mm difference from $\mu_0 = 3000$), the power was about 0.36 with $n = 25$ and about 0.72 with $n = 60$. For detecting $\mu_1 = 3003$ (a 3 mm difference), the power was about 0.998 with $n = 25$ and about 1.00 with $n = 60$. What does this pattern illustrate about the relationship between sample size, effect size, and power?

1) Increasing the sample size has the largest relative impact on power when the true effect size is small (harder to detect); for large effect sizes, power is already close to 1 even with the smaller sample, leaving little room for improvement.
2) Sample size and effect size affect power independently and additively; doubling n always increases power by the same fixed amount, regardless of effect size.

3) Since power increased with n for the 1 mm case, it must also *decrease* with n for the 3 mm case, since power is bounded above by 1.
4) The pattern shows a coding error; power should always be exactly 0.80 for a well-designed study, regardless of n or effect size.
5) Power depends only on the effect size, not on the sample size — the observed changes with n must be due to random simulation noise.

|||| Solution

Correct answer: **1)**. For the 1 mm effect, going from $n = 25$ to $n = 60$ nearly doubles power ($0.36 \rightarrow 0.72$); for the 3 mm effect, power was already essentially at its ceiling of 1 with $n = 25$ ($0.998 \rightarrow 1.00$), so there is little room left to improve. This diminishing-returns pattern — extra sample size matters most when an effect is hard to detect — is a general feature of power curves, not a coding error or noise. Option 3) misunderstands that power can plateau near 1 without needing to "compensate" by decreasing elsewhere.

3.11 Concrete items 2

|||| Exercise 3.11 Concrete items 2

A construction company receives concrete items for a construction. The lengths of the items are assumed reasonably normally distributed, with mean $\mu = 3000$ mm and standard deviation $\sigma = 3$ mm.

a) In a construction process, 5 concrete items are joined together to a single construction with a total length equal to the sum of the 5 concrete items. It is very important that the length of this joined construction is within 15 m plus/minus 2 cm. How often will it happen that such a construction will be more than 2 cm away from the 15 m target?

   (*Use Python to compute probabilities from a relevant normal distribution*)

|||| Solution

The lengths of the joined constructions are subject to some randomness given they are assembled from 5 individual pieces with random lengths (which follow a normal distribution as described). The total joined length is a sum of 5 individual (but random) lengths. Therefore the length of the joined construction is given by $Y = \sum_{i=1}^{5} X_i$, where X_i are random variables describing the lengths of the individual pieces.

Using the rules for linear combinations of independent normals, we compute the mean and variance of Y:

$$E[Y] = \sum_{i=1}^{5} E[X_i] = 3000 + 3000 + 3000 + 3000 + 3000 = 15000,$$

$$V[Y] = \sum_{i=1}^{5} 1^2 \cdot V[X_i] = 3^2 + 3^2 + 3^2 + 3^2 + 3^2 = 45$$

Since Y is a linear combination of normally distributed stochastic variables, Y will also follow a normal distributions. Hence we have: $\mathrm{Y} \sim N(15000, 45)$.

Now we want to calculate the probability that Y deviates 2 centimeters (20 mm) or more from the target of 15 meters (15000 mm), i.e., $P(Y > 15020) + P(Y < 14980)$. Since we know that Y follows a normal distribution we have:

$$P(Y > 15020) = 1 - F(15020)$$
$$P(Y < 14980) = F(14980)$$

where $F(Y)$ is the cdf of Y. We use Python to compute these probabilities:

```
print(1 - stats.norm.cdf(15020, 15000, np.sqrt(45)))
0.0014345

print(stats.norm.cdf(14980, 15000, np.sqrt(45)))
0.0014345
```

So the total probability is $0.0014345 + 0.0014345 = 0.00287$. To conclude: the joined construction will be more than 2 cm away from the 15 m target in roughly 0.3% of cases.

The construction company wants to check that the individual concrete items do in fact have an average length of 3000 mm and a standard deviation of 3 mm (still assuming that the lengths follow a normal distribution). The company samples 9 items from a delivery which are then measured for control. The following measurements (in mm) are found:

3003	3005	2997	3006	2999	2998	3007	3005	3001

b) Compute the following three statistics: the sample mean, the sample standard deviation and the standard error of the mean (*you may use Python for calculations*).
Explain the difference between the sample standard deviation and the standard error of the mean.

|||| Solution

```
x = np.array([3003, 3005, 2997, 3006, 2999, 2998, 3007,
    3005, 3001])
print(x.mean(), x.std(ddof=1), x.std(ddof=1)/np.sqrt(9))
3002.33 3.708 1.236
```

From the data we get the sample mean and sample standard deviation

$$\bar{x} = 3002.33 \text{ mm} \quad \text{and} \quad s = 3.708 \text{ mm},$$

and the standard error of the mean

$$SE_{\bar{x}} = \frac{3.708}{\sqrt{9}} = 1.236.$$

The **sample standard deviation** describes the variation between individual concrete items. The **standard error of the mean** instead describes the expected variation of the sample mean ($\bar{x}$) when calculated from samples of the same size ($n = 9$).

c) Find the 95% confidence interval for the mean μ.

(*Use Python to get the needed quantile(s) from a relevant probability distribution*)

|||| Solution

We use Method 3.9.

For the calculation we need $t_{1-\alpha/2}$ from a t-distribution with 8 degrees of freedom ($\nu = n - 1 = 8$).

Since we are asked to compute the 95%-confidence interval we need the $t_{0.975}$-quantile (you can convince yourself of this by drawing the pdf of a t-distribution and sketch the interval (centered about the mean) corresponding to 95% of the total probability mass (area under the pdf-curve) - this will leave 2.5% of the total probability mass on either side of the interval).

We use Python to find the value of the $t_{0.975}$-quantile:

```
print(stats.t.ppf(0.975, df=8))
2.306
```

We can also write: $t_{0.975,\nu=8} = 2.306$ (to make it more clear which distribution is used).

The 95% confidence interval for the mean is given as:

$$3002.33 \pm 2.306 \cdot \frac{3.708}{\sqrt{9}} \iff [2999.5;\ 3005.2].$$

d) Find the 99% confidence interval for μ. Compare with the 95% one from above and explain why it is smaller/larger.

(*Use Python to get the needed quantile(s) from a relevant probability distribution*)

|||| Solution

The needed quantile is $t_{0.995,\,\nu=8}$, and we find the value using Python:

```
print(stats.t.ppf(0.995, df=8))
3.355
```

The 99% confidence interval for the mean is given as:

$$3002.33 \pm 3.355 \cdot \frac{3.708}{\sqrt{9}} \iff [2998.2;\; 3006.5].$$

The 99% interval is wider: wanting to be *more confident* of capturing the true mean μ requires a *broader interval*.

e) Find the 95% confidence intervals for the variance σ^2 and the standard deviation σ.

(*Use Python to get the needed quantile(s) from a relevant probability distribution*)

|||| Solution

We use Method 3.19.

For the calculation we need $\chi^2_{\alpha/2}$ and $\chi^2_{1-\alpha/2}$ from a χ^2-distribution with 8 degrees of freedom ($\nu = n - 1 = 8$).

Since we are asked to compute the 95%-confidence interval we need the $\chi^2_{0.025}$-quantile and the $\chi^2_{0.975}$-quantile (you can convince yourself of this by drawing the pdf of a χ^2-distribution and sketch the interval corresponding to 95% of the total probability mass (area under the pdf-curve), leaving 2.5% of the total probability mass on the left side of the interval and 2.5% of the total probability mass on the right side of the interval).

We use Python to find the values of the quantiles: $\chi^2_{0.025,\,\nu=8}$ and $\chi^2_{0.975,\,\nu=8}$:

```
print(stats.chi2.ppf(0.025, df=8))
2.180
```

```
print(stats.chi2.ppf(0.975, df=8))
17.535
```

Using $(n-1)s^2 = 8 \times 3.708^2 = 110.0$:

$$\left[\frac{8 \times 13.75}{17.535};\ \frac{8 \times 13.75}{2.180}\right] = [6.27;\ 50.46].$$

For the standard deviation:

$$[\sqrt{6.27};\ \sqrt{50.46}] = [2.50;\ 7.10].$$

f) Based on the 95% confidence intervals, what should the construction company conclude about the delivered concrete items? Can they trust that the lengths of the individual items have mean 3000 mm and standard deviation 3 mm?

|||| Solution

The company cannot be absolutely sure about the exact true values of μ and σ, but since the 95% CI for μ includes 3000 and the 95% CI for σ includes 3, these values are entirely consistent with the data, even though the sample estimates $\bar{x}$ and s are not exactly equal to 3000 and 3 respectively.

g) Suppose instead the construction company samples $n = 100$ concrete items (instead of 9), and obtains sample estimates similar to before: $\bar{x} = 3002.3$ mm and $s = 3.7$ mm. Based on 95% confidence intervals, what conclusion would the company now reach?

(*Use Python to get the needed quantile(s) from a relevant probability distribution*)

|||| Solution

In this situation the number of observations in the sample is *larger* ($n = 100$), which makes the estimates *more reliable* - this means we would expect *narrower confidence intervals*!

When performing the calculations, the only difference is the degrees of freedom used in the probability distributions: $\nu = 100 - 1 = 99$ (for both the relevant t-distribution and the relevant χ^2-distribution).

For the 95% confidence interval for the mean we need the quantile, $t_{0.975,\nu=99} = 1.984$:

```
print(stats.t.ppf(0.975, df=99))
1.984
```

(*notice the value* $t_{0.975,\nu=99} = 1.984$ *is very close to the corresponding value from a standard normal distribution* $z_{0.975} = 1.96$, *because the degrees of freedom,* ν, *is quite large in this case*).

We now compute the interval:

$$3002.3 \pm 1.984 \cdot \frac{3.7}{\sqrt{100}} \iff [3001.57;\ 3003.03].$$

For the 95% confidence interval for the variance (and standard deviation) we need the quantiles, $\chi^2_{0.025,\nu=99}$ and $\chi^2_{0.025,\nu=99}$:

```
print(stats.chi2.ppf(0.025, df=99))
73.36

print(stats.chi2.ppf(0.975, df=99))
128.42
```

Using $(n-1)s^2 = 99 \times 3.7^2 = 1355.31$:

$$\left[\frac{99 \times 3.7^2}{128.42};\ \frac{99 \times 3.7^2}{73.36}\right] = [10.55;\ 18.47].$$

For the standard deviation:

$$[\sqrt{10.55};\ \sqrt{18.47}] = [3.25;\ 4.30].$$

This gives $\mu \in [3001.6; 3003.0]$ and $\sigma \in [3.25; 4.30]$. Neither interval contains the stated values 3000 mm and 3 mm, even though the point estimates ($\bar{x} = 3002.3$, $s = 3.7$) are almost identical to the $n = 9$ sample above, so the company would now suspect (and likely conclude) that the true population mean and standard deviation are both somewhat larger than expected. The extra precision from the larger sample reveals a deviation that the small sample above could not.

3.12 Concrete items 2 (hypothesis testing)

Exercise 3.12 Concrete items 2 (hypothesis testing)

This is a continuation of Exercise 11:

A construction company receives concrete items for a construction. The lengths of the items are assumed reasonably normally distributed, with mean $\mu = 3000$ mm and standard deviation $\sigma = 3$ mm.

The company samples 9 items from a delivery which are then measured for control. The following measurements (in mm) are found:

3003	3005	2997	3006	2999	2998	3007	3005	3001

a) To investigate whether the requirement on the mean is fulfilled (with $\alpha = 5\%$), test the following hypothesis:

$$H_0 : \mu = 3000$$
$$H_1 : \mu \neq 3000.$$

Or similarly asked: what is the evidence against the null hypothesis?

(*Use Python to compute quantiles and probabilities from relevant probability distribution(s).*)

Solution

From Exercise 11, we already know $\bar{x} = 3002.33$ and $s = 3.708$:

```
x = np.array([3003, 3005, 2997, 3006, 2999, 2998, 3007,
    3005, 3001])
print(x.mean(), x.std(ddof=1))
3002.33 3.708
```

We compute t_{obs} (Method 3.23):

$$t_{\text{obs}} = \frac{3002.33 - 3000}{3.708/\sqrt{9}} = 1.885.$$

The p-value (using a t-distribution with $\nu = n - 1 = 8$ degrees of freedom) is then calculated as (Method 3.23):

$$p\text{-value} = 2 \cdot P(T > 1.885) = 2 \cdot P(T < -1.885) = 0.096$$

where the actual value was computed using Python:

```
print(2*stats.t.cdf(-1.885, df=8))
0.0961
```

(notice the use of symmetry: $P(T > 1.885) = P(T < -1.885)$)

Conclusion: There is weak evidence against the null hypothesis (see Table 3.1); at $\alpha = 0.05$ we do not reject H_0.

The test statistic and p-value could also be calculated in Python (after having saved the raw data in the variable x) in the following way:

```
t_obs, p_val = stats.ttest_1samp(x, popmean=3000)
print(t_obs, p_val)

1.887 0.0957
```

b) At significance level $\alpha = 0.01$, what are the critical values for t_{obs}, and what is their interpretation?

(*Use Python to compute quantiles and probabilities from relevant probability distribution(s).*)

|||| Solution

The critical values are $\pm t_{0.995,\nu=8} = \pm 3.355$, as found using Python:

```
print(stats.t.ppf(0.995, df=8))
3.355
```

At $\alpha = 0.01$, H_0 is rejected only if t_{obs} falls outside the range $[-3.355, 3.355]$. Since the observed $t_{\text{obs}} = 1.885$ lies well within this interval, H_0 is not rejected.

c) At significance level $\alpha = 0.05$, what are the critical values for t_{obs}? And what is the interpretation of these critical values (compare also with the values found in the previous question)?

(*Use Python to compute quantiles and probabilities from relevant probability distribution(s).*)

|||| Solution

The critical values are $\pm t_{0.975,\nu=8} = \pm 2.306$, as found using Python:

```
print(stats.t.ppf(0.975, df=8))
2.306
```

Since $t_{\text{obs}} = 1.885$ is well within the interval $[-2.306, 2.306]$, H_0 is not rejected at $\alpha = 0.05$.

A smaller α corresponds to a larger *acceptance region* for t_{obs} (i.e., larger critical values): choosing $\alpha = 0.01$ demands stronger evidence against the null hypothesis H_0 than $\alpha = 0.05$ does.
Here "*stronger evidence against the null hypothesis*" corresponds to obtaining a sample average, $\bar{x}$, which is further away from the null hypothesis value μ_0 (which would lead to a larger t_{obs} - you can convince yourself of this by inspecting the formula for t_{obs}).

d) As mentioned in the beginning of the exercise the lengths of the concrete items are assumed reasonably normally distributed.
Is is assumption important for the results calculated above? Why?

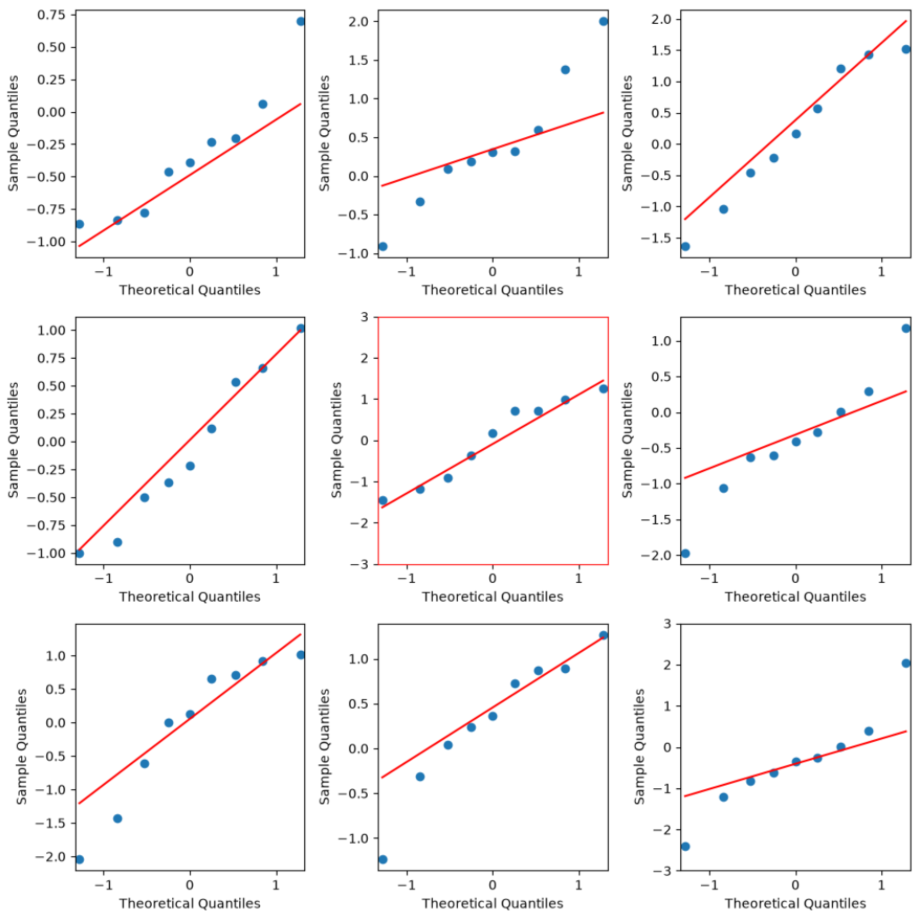
|||| Solution

Yes, the normality assumption is important because the hypothesis test is based on the assumption that the estimator, namely the sample mean $\bar{x}$, follows a normal distribution. When the population of concrete item lengths is normally distributed, the sample mean is also normally distributed, and the test statistic follows a t-distribution. This is what allows us to calculate valid p-values and confidence intervals. Since the sample size is small ($n = 9 \leq 30$), this assumption is important for the validity of the results.

e) The construction company now investigates visually whether the data appears to be coming from a normal distribution (as assumed until now).

They produce a QQ-plot of the data and compare with QQ-plots of *8 simulated data samples* (all of size $n = 9$) sampled from normal distributions with $\mu = \bar{x}$ (mean equal to the sample mean) and $\sigma = s$ (standard deviation equal to the sample standard deviation).

The QQ-plots looked like this (were the QQ-plot of the true data was plotted in the middle):



Does the plot give rise to concerns about the normal assumption?

Solution

No, the plot does not give rise to concerns about the normal assumption. The true data seems to follow the straight line equally well as the simulated data.

The code for constructing the plot is given here (if you want to try it yourself):

```
# Enter data into variable x:
x = np.array([3003,3005,2997,3006,2999,2998,3007,3005,3001])

# Generate 9 plots
fig, axs = plt.subplots(3, 3, figsize=(10, 10))

# Create 9 QQ plots of random normal samples (simulated
   normal data)
for ax in axs.flat:
    sm.qqplot(stats.norm.rvs(size=9), line="q", ax=ax)
ax.set_ylim([-3, 3])

# Replace middle plot with real data
axs[1,1].clear()
sm.qqplot((x-x.mean())/x.std(ddof=1), line="q", ax=axs[1,1])
axs[1,1].set_ylim([-3, 3])

# Highlight middle plot with red spines
plt.setp(axs[1,1].spines.values(), color="red")

# Generate the plot
plt.tight_layout()
plt.show()
```