Applied Statistics for Life Sciences

Dmitri D. Pervouchine

Centre de Regulació Genòmica

Module 3: Statistical Inference, Part I.

Contents

"There are three kinds of lies: lies, damned lies, and statistics."

Benjamin Disraeli

Our objective is to make exact, yet probabilistic statements about population based on the incomplete information (i.e., sample) that was actually observed. In statistics, we don't prove or disprove; we simply find evidence for or against certain hypotheses.

Hypothesis Testing

- H_0 — null hypothesis
- H_a — alternative hypothesis

In a court, if a jury rejects the presumption of innocence, the defendant is pronounced guilty, i.e.

- H_0 — the person is not guilty
- H_a — the person is guilty

During medical check-up

- H_0 — the patient is sick
- H_a — the patient is not sick

Type I/II error

- Type I error (α) is the error of rejecting a null hypothesis when it is actually true
- Type II error (β) is the error of failing to reject a null hypothesis when it is in fact false
- $\alpha \leftrightarrow \beta$ if $H_0 \leftrightarrow H_a$

Note that we neither prove nor disprove H_0

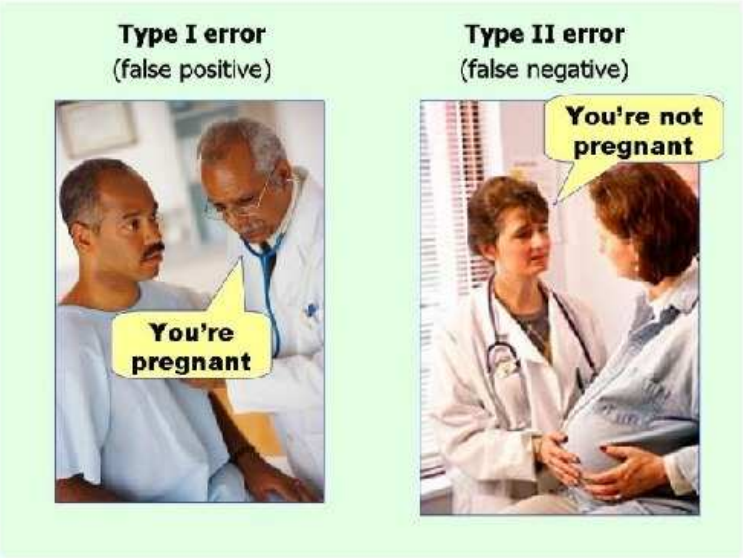
We believe or not believe in it!

Decision rule

- Assume we get many samples
- We set up a decision rule which rejects or accepts the hull hypothesis for each sample
- Sometimes we will commit Type I error
- Sometimes we will commit Type II error
- (Of course many times we will be correct!)
- **Decision rule comes separately from the set of hypotheses!**

False Positives and False Negatives

Actual condition	Test shows	
	"not pregnant"	"pregnant"
H_0 : Not pregnant	True Negative	False Positive **Type I error**
H_a : Pregnant	False Negative **Type II error**	True Positive



Confusion matrix

Actual condition	Test shows		Σ
	"not pregnant"	"pregnant"	
H_0 : Not pregnant	TN	FP	TN + FP
H_a : Pregnant	FN	TP	FN + TP

Actual condition	Test shows		Σ
	"not pregnant"	"pregnant"	
H_0 : Not pregnant	$\frac{TN}{TN+FP} =$ True Negative Rate = $1-\alpha =$ Specificity	$\frac{FP}{TN+FP} = \alpha =$ False Positive Rate	100%
H_a : Pregnant	$\frac{FN}{FN+TP} = \beta =$ False Negative Rate	$\frac{TP}{FN+TP} = 1-\beta =$ True Positive Rate = Sensitivity	100%

Confusion matrix

Actual condition	Test shows	
	"not pregnant"	"pregnant"
H_0 : Not pregnant	TN	FP
H_a : Pregnant	FN	TP
Σ	TN + FN	FP + TP

Actual condition	Test shows	
	"not pregnant"	"pregnant"
H_0 : Not pregnant	$\frac{TN}{TN+FN} =$ Negative Predictive Value	$\frac{FP}{FP+TP} =$ False Discovery Rate
H_a : Pregnant	$\frac{FN}{TN+FN}$	$\frac{TP}{FP+TP} =$ Positive Predictive Value = Precision
Σ	100%	100%

"Statistics is a form of quantitative discourse."

Discourse

a: formal and orderly and usually extended expression of thought on a subject

b: connected speech or writing

c: a *linguistic* unit (as a conversation or a story) larger than a sentence

Meriam Webster Dictionary

Odds ratio

- Likelihood ratio positive $= \frac{\text{sensitivity}}{1-\text{specificity}}$
- Likelihood ratio negative $= \frac{\text{1-sensitivity}}{\text{specificity}}$
- Odds ratio $= \frac{\text{Likelihood ratio positive}}{\text{Likelihood ratio negative}} = \frac{TP/FP}{FN/TN}$
- $\alpha =$ Significance level
- $1 - \alpha =$ Confidence level
- $1 - \beta =$ Power

Patients with bowel cancer are screened with fecal occult blood screen test

	Test comes out		
Cancer	Negative	Positive	
No	TN = 1820	FP = 180	2000
Yes	FN = 10	TP = 20	30
	1830	200	2030

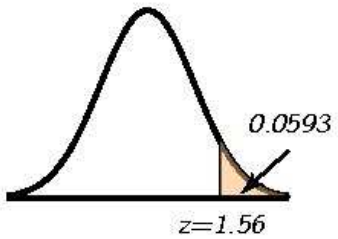
- False positive rate $\alpha = FP/(FP + TN) = 180/(180 + 1820) = 9\%$
- False negative rate $\beta = FN/(TP + FN) = 10/(20 + 10) = 33\%$
- Power = sensitivity $= 1 - \beta = 67\%$
- Specificity $= 1 - \alpha = 91\%$
- Likelihood ratio positive $= \frac{sensitivity}{1-specificity} = \frac{0.67}{1-0.91} = 7.4$
- Likelihood ratio negative $= \frac{1-sensitivity}{specificity} = \frac{1-0.67}{0.91} = 0.37$
- LR$=\frac{7.4}{0.37} = 20$, strong predictive power of the test

Problem 3.1

A patient claims that he consumes only 2000 calories per day, but a dietician suspects that the actual figure is higher. The dietician plans to check his food intake for 30 days and will reject the patient's claim if the 30-day-mean is more than 2100 calories. If the standard deviation (in calories per day) is 350, what is the probability that the dietician will mistakenly reject a patient's true claim?

Solution

- $H_0 : \mu = 2000$
- $H_a : \mu > 2000$
- H_0 *is rejected whenever we get a sample with* $\overline{X} > 2100$
- $\mathrm{P}(\overline{X} > 2100) = \mathrm{P}(Z > \frac{2100-2000}{350/\sqrt{30}}) = \mathrm{P}(Z > 1.56) = 0.0593$

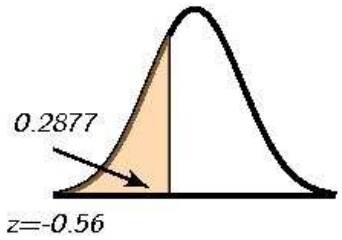


Problem 3.2

City planners wish to test the claim that shoppers park for an average of only 47 minutes in the downtown area. The planners have decided to tabulate parking durations for 225 shoppers and to reject the claim if the sample mean exceeds 50 minutes. If the claim is wrong and the true mean is 51 minutes, what is the probability that the random sample will lead to a mistaken failure to reject the claim? Assume that the standard deviation in parking durations is 27 minutes.

Solution

- $H_0 : \mu = 47$
- $H_a : \mu > 47$
- H_0 *is rejected whenever we get a sample with* $\overline{X} > 50$
- $\mathrm{P}(\overline{X} < 50) = \mathrm{P}(Z < \frac{50-51}{27/\sqrt{225}}) = \mathrm{P}(Z < -0.56) = 0.2877$



P-value is the probability of obtaining a result at least as extreme as the one that was actually observed, given that the null hypothesis is true.

In other words, if the null hypothesis were true, what would be the probability to get the sample that we have got?

P-value

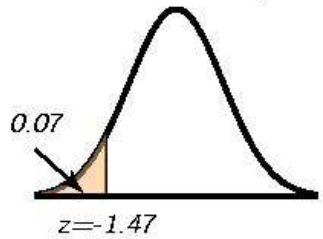
- P-value is a function of a sample
- α is a function of a decision rule
- Reject H_0 if P-value $< \alpha$
- Small P-value indicates that what you see would have been very unusual if H_0 were true

Problem 3.3

A coffee machine is supposed to deliver 8 ounces of coffee per cup. A random sample of 50 cups has the mean of 7.75 ounces and standard deviation of 1.2 ounces. Is there a reason to believe that the coffee machine is not operating as it should?

Solution

- $H_0 : \mu = 8$
- $H_a : \mu < 8$
- $\sigma(\overline{X}) = \frac{\sigma}{\sqrt{n}} = \frac{1.2}{\sqrt{50}} = 0.1697$
- P-value $= \mathrm{P}(\overline{X} < 7.75) = \mathrm{P}(Z < \frac{7.75-8}{0.1697}) = \mathrm{P}(Z < -1.47) = 0.07$



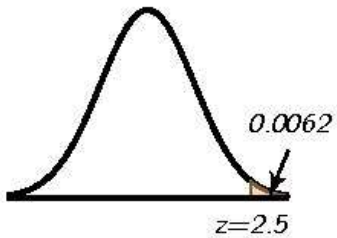
- P-value $= 0.07 > \alpha = 0.05$, *i.e., there is no sufficient evidence against the 8 ounces claim. That is, it is not too surprising to obtain a sample with* $\overline{X} = 7.75$ *if the actual* μ *were 8 oz.*

Problem 3.4

A service station advertises that its mechanics can change a muffler in only 15 minutes. A consumers group doubts this claim and runs a hypothesis test using 49 cars needing new mufflers. In this sample the mean changing time is 16.25 minutes with a standard deviation of 3.5 minutes. Is this a strong evidence against the 15 minute claim?

Solution

- $H_0 : \mu = 15$
- $H_a : \mu > 15$
- P-value $= \mathrm{P}(\overline{X} > 16.25) = \mathrm{P}(Z > \frac{16.25-15}{3.5/\sqrt{49}}) = \mathrm{P}(Z > 2.5) = 0.0062$



- P-value $= 0.0062 < \alpha = 0.05$, *i.e., there is sufficient evidence against the 15 minute claim. That is, it would be too surprising to obtain a sample with* $\overline{X} > 16.25$ *if the actual* μ *were 15 min.*

Estimators

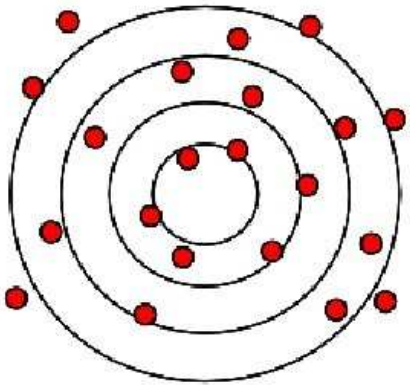
An *estimator* is a function of the observable sample data that is used to estimate an unknown population parameter

- $\overline{X}$ is an estimator for μ
- s is an estimator for σ
- $\hat{p}$ is an estimator for p

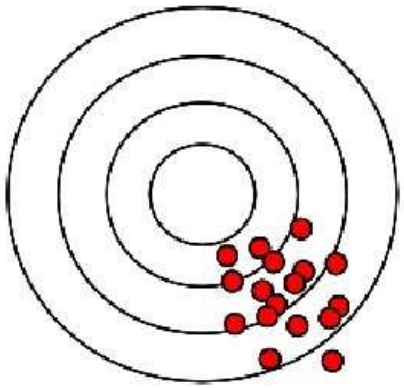
Unbiased and Effective Estimators

- Let θ be the unknown parameter
- Let $\hat{\theta}_n$ be an estimator
- $\hat{\theta}_n$ is *unbiased* if $\mathrm{E}(\hat{\theta}_n) = \theta$
- $\hat{\theta}_n$ is *effective* if $\lim_{n \to \infty} \mathrm{Var}(\hat{\theta}_n) = 0$

Unbiased vs. Effective Estimators



Unbiased but ineffective



Effective but biased

We are looking for unbiased and effective estimators

Mean Square Error

- Bias

$$bias(\hat{\theta}_n) = \mathsf{E}(\hat{\theta}_n) - \theta$$

- Variance

$$\mathsf{Var}(\hat{\theta}_n) = \mathsf{E}\left(\hat{\theta}_n - \mathsf{E}(\hat{\theta}_n)\right)^2$$

- Mean Square Error

$$MSE(\hat{\theta}_n) = \mathsf{E}\left(\hat{\theta}_n - \theta\right)^2 = \mathsf{Var}(\hat{\theta}_n) + bias(\hat{\theta}_n)^2$$

Sample mean is an unbiased effective estimator

- $\mathsf{E}(\overline{X}) = \mathsf{E}\left(\frac{X_1+X_2+\ldots+X_n}{n}\right) = \frac{1}{n}\mathsf{E}(X_1 + X_2 + \ldots + X_n) = \frac{1}{n}(\mu + \mu + \ldots + \mu) = \frac{n\mu}{n} = \mu$

- $\mathsf{Var}(\overline{X}) = \mathsf{Var}\left(\frac{X_1+X_2+\ldots+X_n}{n}\right) = \frac{1}{n^2}\mathsf{Var}(X_1 + X_2 + \ldots + X_n) = \frac{1}{n^2}(\sigma^2 + \sigma^2 + \ldots + \sigma^2) = \frac{n\sigma^2}{n^2} = \sigma^2/n \to 0$

Sample variance is an unbiased estimator

- $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2 = \frac{1}{n-1}\left(\sum_{i=1}^{n} X_i^2 - n\overline{X}^2\right)$
- $\mathrm{E}(X_i) = \mu, \quad \mathrm{Var}(X_i) = \sigma^2, \quad \mathrm{E}(X_i^2) = \sigma^2 + \mu^2$
- $\mathrm{E}(\overline{X}) = \mu, \quad \mathrm{Var}(\overline{X}) = \sigma^2/n, \quad \mathrm{E}(\overline{X}^2) = \sigma^2/n + \mu^2$
- $\mathrm{E}(s^2) = \frac{1}{n-1}\left(n(\sigma^2 + \mu^2) - n(\sigma^2/n + \mu^2)\right) = \frac{n-1}{n-1}\sigma^2$
- Indeed, sample variance is an unbiased estimator for population variance.

Problem 3.5

A box contains 70 black and 30 white balls. Ten balls are chosen at random and two estimators for the proportion of black balls are considered

$$\hat{p}_1 = \frac{\textit{the number of black balls}}{n},$$

$$\hat{p}_2 = \frac{\textit{the number of black balls}+2}{n+2},$$

where n=10. Which estimator is more effective? (i.e., has smaller MSE?)

Solution

- $X \sim Bi(n = 10, p = 0.7)$
- $\hat{p}_1 = \frac{X}{n}$, $\mathrm{E}(\hat{p}_1) = \frac{np}{n} = 0.7$, $bias(\hat{p}_1) = 0.7 - 0.7 = 0$,
- $\mathrm{Var}(\hat{p}_1) = \frac{np(1-p)}{n^2} = \frac{0.3 \cdot 0.7}{10} = 0.021$, $MSE(\hat{p}_1) = 0.021 + 0^2 = 0.021$
- $\hat{p}_2 = \frac{X+2}{n+2}$, $\mathrm{E}(\hat{p}_1) = \frac{np+2}{n+2} = 0.75$, $bias(\hat{p}_2) = 0.75 - 0.7 = 0.05$,
- $\mathrm{Var}(\hat{p}_2) = \frac{np(1-p)}{(n+2)^2} = \frac{10 \cdot 0.3 \cdot 0.7}{12^2} = 0.0146$, $MSE(\hat{p}_2) = 0.0146 + 0.05^2 = 0.0171$, $\hat{p}_2$ **is more effective!**

Standard Error

Standard error is the standard deviation of the estimator

$$\mathrm{SE}(\overline{X}) = \frac{\sigma}{\sqrt{n}}$$

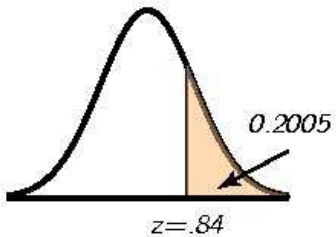
$$\mathrm{SE}(\hat{p}) = \sqrt{\frac{p(1-p)}{n}}$$

Problem 3.6

A local restaurant owner claims that only 15% of visiting tourists stay for more than 2 days. A chamber of commerce volunteer is sure that the real percentage is higher. He plans to survey 100 tourists and intends to speak up if at least 18 of the tourists stay longer than 2 days. What is the probability of mistakenly rejecting the restaurant owner's claim if it is true?

Solution

- $H_0 : p = 0.15$
- $H_a : p > 0.15$
- *Reject* H_0 *whenever we get a sample with* $\hat{p} > 0.18$
- $\text{P-value} = \text{P}(\hat{p} > 0.18) = \text{P}(Z > \frac{0.18-0.15}{\sqrt{\frac{0.15 \cdot 0.85}{100}}}) = \text{P}(Z > 0.84) = 0.2005$



Two-sample mean

Consider two **independent** samples $X_1, X_2, \ldots, X_n$ and $Y_1, Y_2, \ldots, Y_m$ from two populations with population means μ_1 and μ_2 and population variances σ_1^2 and σ_2^2, respectively.

$$\mathrm{SE}(\overline{X} - \overline{Y}) = \sqrt{\frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}},$$

$$\mathrm{SE}(\overline{X} - \overline{Y}) = \sigma\sqrt{\frac{1}{n} + \frac{1}{m}}, \text{ if } \sigma_1 = \sigma_2.$$

Two-sample proportion

Two independent sample proportions

$$\mathrm{SE}(\hat{p}_1 - \hat{p}_2) = \sqrt{\frac{p_1(1-p_1)}{n} + \frac{p_2(1-p_2)}{m}},$$

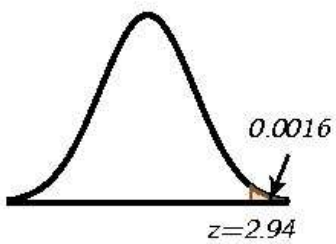
$$\mathrm{SE}(\hat{p}_1 - \hat{p}_2) = \sqrt{p(1-p)}\sqrt{\frac{1}{n} + \frac{1}{m}}, \text{ if } p_1 = p_2.$$

Problem 3.7

A historian believes that the average height of soldiers in World War II was greater than that of soldiers in World War I. She examines a random sample of records of 100 men in each war and notes standard deviations of 2.5 and 2.3 inches in World War I and World War II, respectively. If the average height from the sample of World War II soldiers is 1 inch greater than that from the sample of World War I soldiers, what conclusion is justified from a two-sample hypothesis test where $H_0 : \mu_1 = \mu_2$ *vs.* $H_a : \mu_1 < \mu_2$*?*

Solution

- $H_0 : \mu_2 - \mu_1 = 0$
- $H_a : \mu_2 - \mu_1 > 0$
- *Estimator* $= \overline{Y} - \overline{X} =$ *difference of sample means*
- P-value $= \mathrm{P}(\overline{Y} - \overline{X} > 1) = \mathrm{P}(Z > \frac{1-0}{\sqrt{\frac{2.5^2}{100} + \frac{2.3^2}{100}}}) = \mathrm{P}(Z > 2.94) = 0.0016$



There is enough evidence at 5% significance level that the average height of WW-II solders was greater than that of WW-I solders.

Confidence Intervals

$$\text{Parameter} = \text{Estimate} \pm \text{Critical} * \text{SE}$$

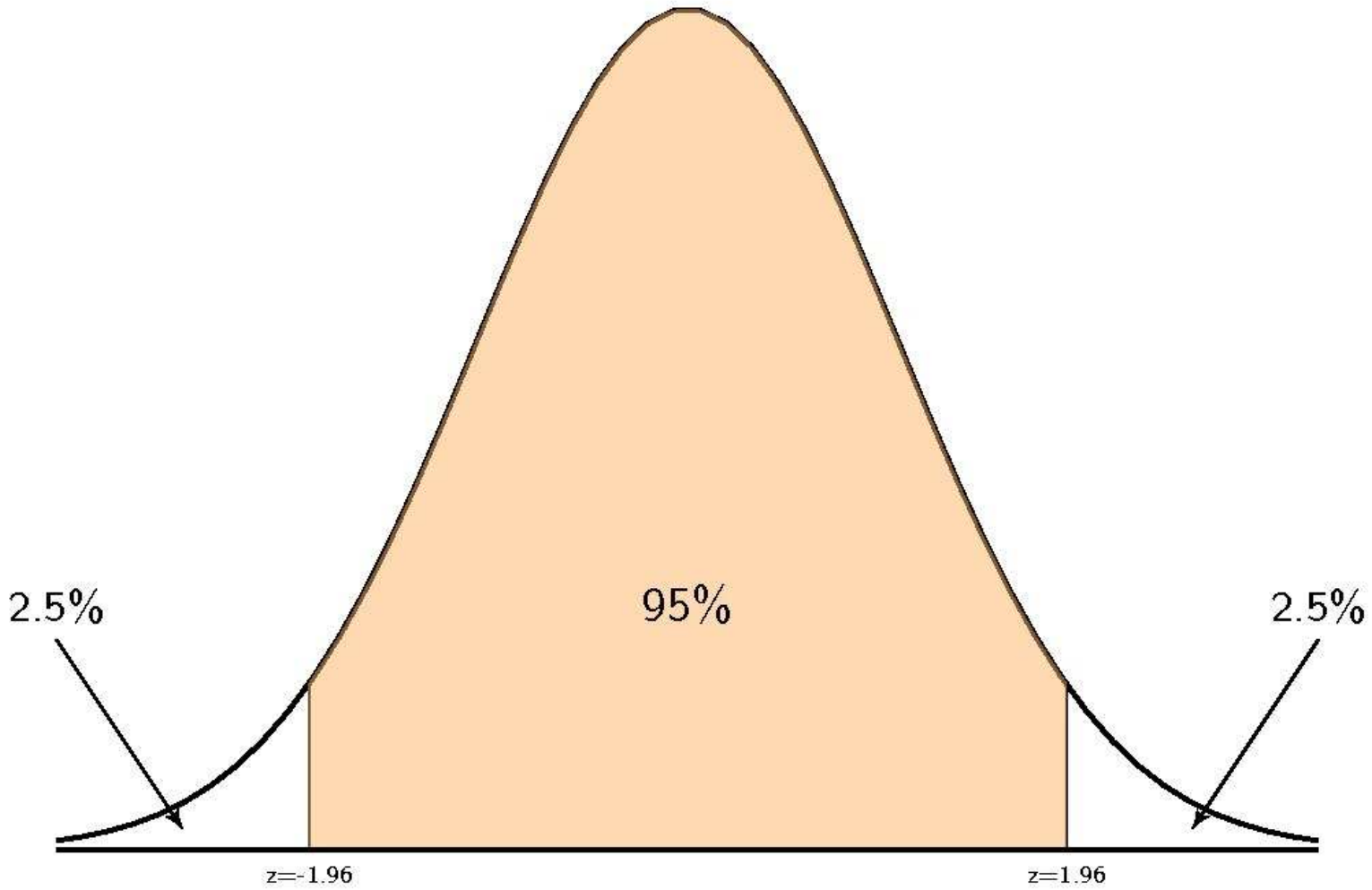
$$\text{SE} = \text{Standard Error}$$

$$\text{Critical} = \text{critical Value}$$

$$\mu = \overline{X} \pm z_{\alpha/2} \cdot \text{SE}, \text{where } \text{SE} = \frac{\sigma}{\sqrt{n}},$$

$$p = \hat{p} \pm z_{\alpha/2} \cdot \text{SE}, \text{where } \text{SE} = \sqrt{\frac{p(1-p)}{n}}.$$

Critical value

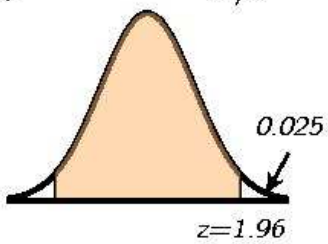


Problem 4.1 (problem 3.1 revisited)

A patient claims that he consumes only 2000 calories per day, but a dietician suspects that the actual figure is higher. The dietician checked his food intake for 30 days and found that the 30-day-mean is more than 2100 calories. What is the 95% confidence interval for the number of calories in patient's diet? Assume standard deviation of 350 calories per day.

Solution

- $\mu = \overline{X} \pm z_{\alpha/2} \cdot \text{SE}$, *where* $\alpha/2 = (1 - 0.95)/2 = 0.025$



- $\mu = 2100 \pm 1.96 \cdot \frac{350}{\sqrt{30}} = 2100 \pm 125 = [1975, 2225]$
- *We are 95% confident that the value of μ is between 1975 and 2225 cal*

Interpretation

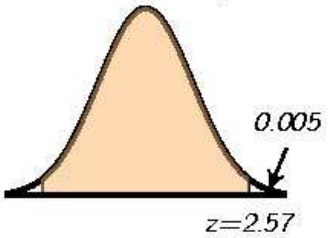
- We are 95% confident that the value of μ is in the confidence interval that we built
- 95% of samples result in confidence intervals which contain the true value of μ
- If we believe that our sample is "typical", i.e., within those 95% of samples, then yes the confidence interval that we built contains the true value of μ
- **Note that $P(\mu \in [a, b]) = 0.95$ is wrong**

Problem 4.2 (problem 3.6 revisited)

A chamber of commerce volunteer is interested in the percentage of visiting tourists staying for more than 2 days in a certain hotel. He surveyed 100 tourists and found that 18 of them stay longer than 2 days. What is the 99% confidence interval for the percentage of visiting tourists who stay for more than 2 days?

Solution

- $p = \hat{p} \pm z_{\alpha/2} \cdot \text{SE}$, *where* $\alpha/2 = (1 - 0.99)/2 = 0.005$



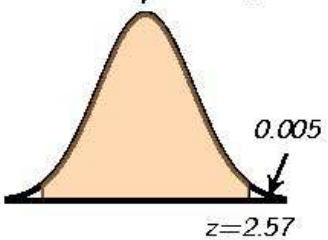
- $p = 0.18 \pm 2.57 \sqrt{\frac{0.18 \cdot 0.82}{100}} = 0.18 \pm 0.09 = [9\%, 27\%]$
- *We are 99% confident that between 9% and 27% of visitors stay for more than 2 days.*
- *Note that we replace p by $\hat{p}$ for the purpose of computing standard error.*

Problem 4.3

In a random sample of 300 high school students, 225 said they managed time effectively, while in a similar sample of 270 college students, only 108 felt they were effective time managers. What is a 99% confidence interval estimate for the difference between the proportions of high school and colleges students who think they manage time effectively?

Solution

- *Estimator* $= \hat{p}_1 - \hat{p}_2 =$ *difference of sample proportions*
- $\hat{p}_1 = \frac{225}{300} = 0.75$, $\hat{p}_2 = \frac{108}{270} = 0.40$,
- $SE = \sqrt{\frac{p_1(1-p_1)}{n} + \frac{p_2(1-p_2)}{m}} = \sqrt{\frac{0.75 \cdot 0.25}{300} + \frac{0.40 \cdot 0.60}{270}} = 0.0389$



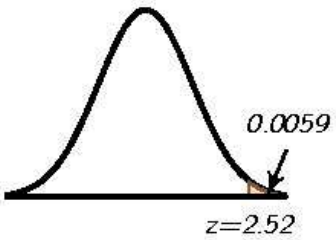
- $p_1 - p_2 = (0.75 - 0.4) \pm 2.57 \cdot 0.0389 = 0.35 \pm 0.10 = [25\%, 45\%]$
- *We are 99% confident that the proportion difference is between 25% and 45%.*

Problem 4.4

A medical researcher believes that taking 1000 milligrams of vitamin C per day will result in fewer colds than a daily intake of 500 milligrams will. In a group of 50 volunteers taking 1000 milligrams per day, the numbers of colds per individual during a winter season averaged 1.8 with a variance of 1.5. Similar data from a group of 60 volunteers taking 500 milligrams per day showed an average of 2.4 with a variance of 1.6. What was the P-value *of this test?*

Solution

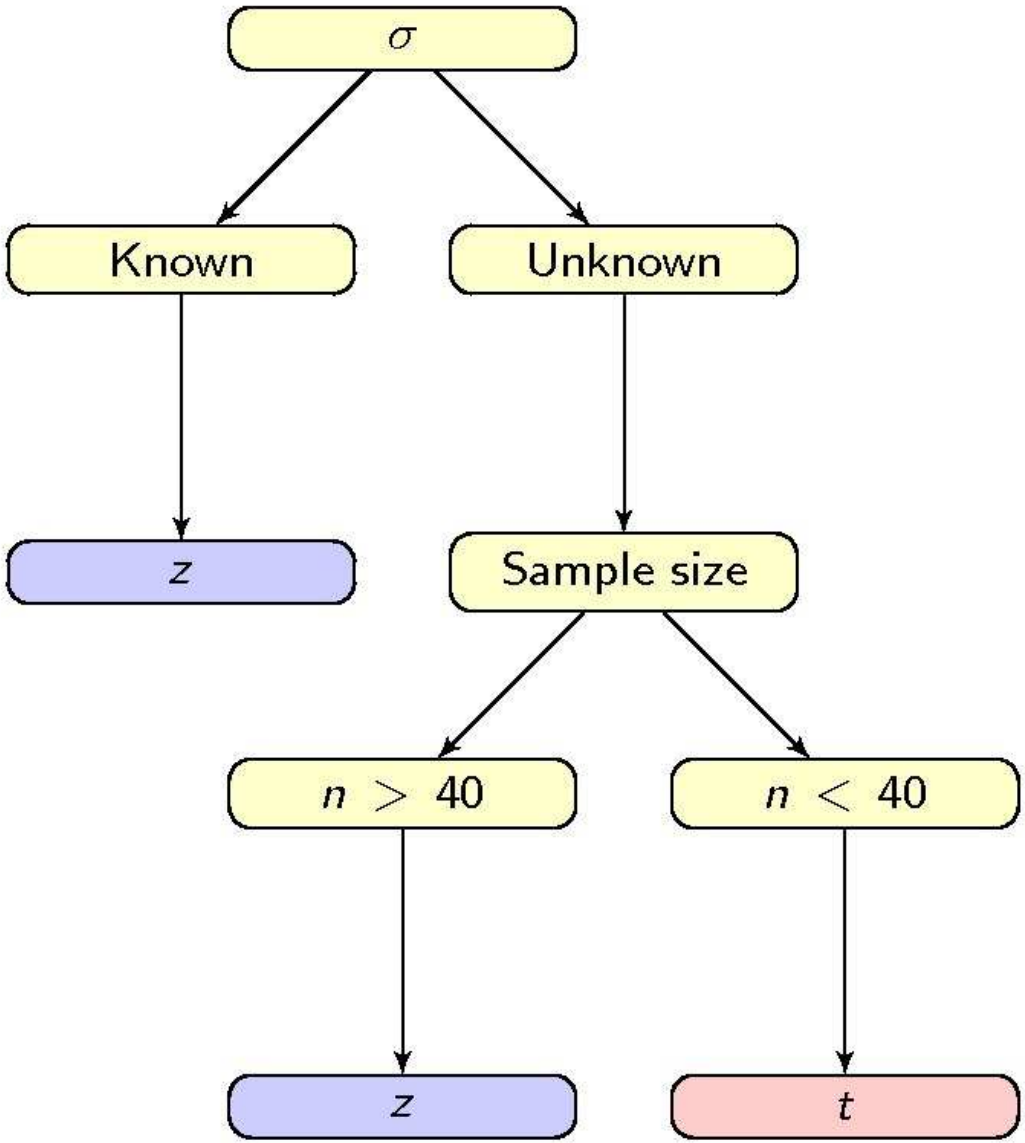
- *Estimator* $= \overline{Y} - \overline{X} =$ *difference of sample means*
- $H_0 : \mu_2 - \mu_1 = 0$
- $H_a : \mu_2 - \mu_1 > 0$
- P-value $= P(\overline{Y} - \overline{X} > 2.4 - 1.8) = P(Z > \frac{0.6-0}{\sqrt{\frac{1.5}{50}+\frac{1.6}{60}}}) = P(Z > 2.52) = 0.0059$



How do we get σ?

- Population standard deviation is usually unknown
- If sample size is large ($n > 40$) then we can assume that the sample standard deviation (s) approximates the population standard deviation (σ) well enough
- If sample size is small then this assumption is no longer valid, i.e., sampling error in the estimation of σ cannot be ignored.

Known vs. unknown σ



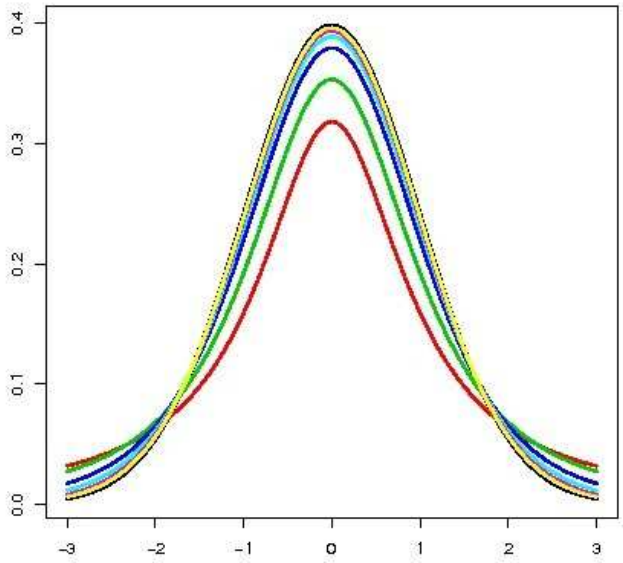
Student t-distribution

$$z = \frac{\overline{X} - \mu}{\sigma / \sqrt{n}}$$

$$t = \frac{\overline{X} - \mu}{s / \sqrt{n}}$$

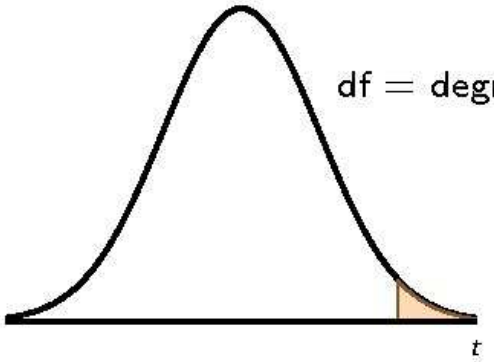
Student t-distribution

- Student t-distribution has one parameter called degrees of freedom



- When the number of degrees of freedom is large, the t-distribution is close to the standard normal distribution

t-distribution table



df = degrees of freedom = sample size - 1

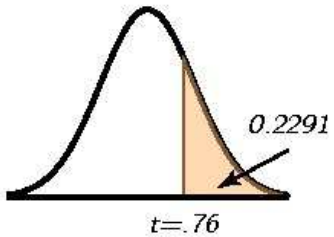
	Tail probability						
df	0.05	0.025	0.01	0.005	0.0025	0.001	0.0005
1	6.3138	12.7065	31.8193	63.6551	127.3447	318.4930	636.0450
2	2.9200	4.3026	6.9646	9.9247	14.0887	22.3276	31.5989
3	2.3534	3.1824	4.5407	5.8408	7.4534	10.2145	12.9242
4	2.1319	2.7764	3.7470	4.6041	5.5976	7.1732	8.6103
5	2.0150	2.5706	3.3650	4.0322	4.7734	5.8934	6.8688
6	1.9432	2.4469	3.1426	3.7074	4.3168	5.2076	5.9589
7	1.8946	2.3646	2.9980	3.4995	4.0294	4.7852	5.4079
8	1.8595	2.3060	2.8965	3.3554	3.8325	4.5008	5.0414
9	1.8331	2.2621	2.8214	3.2498	3.6896	4.2969	4.7809
10	1.8124	2.2282	2.7638	3.1693	3.5814	4.1437	4.5869
11	1.7959	2.2010	2.7181	3.1058	3.4966	4.0247	4.4369
12	1.7823	2.1788	2.6810	3.0545	3.4284	3.9296	4.3178
13	1.7709	2.1604	2.6503	3.0123	3.3725	3.8520	4.2208
14	1.7613	2.1448	2.6245	2.9768	3.3257	3.7874	4.1404
$+\infty$	1.282	1.645	1.960	2.326	2.576	3.091	3.291

Problem 6.1

An article ("Undergraduate Marijuana use and Anger" by Sue Stoner) in a 1988 issue of the Journal of Psychology (Vol. 122, p. 33) reported that in a sample of 17 marijuana users the mean and standard deviation on an anger expression scale were 42.72 and 6.05, respectively. Test whether this result is significantly greater than the established mean of 41.6 for non-users. What assumptions are necessary for the above test to be valid?

Solution

- $H_0 : \mu = 41.6$
- $H_a : \mu > 41.6$
- P-value $= \mathrm{P}(\overline{X} > 42.72) = \mathrm{P}(t(16) > \frac{42.72-41.6}{6.05/\sqrt{17}}) = \mathrm{P}(t(16) > 0.76) = 0.2291$



- *At 5% significance level there is not sufficient evidence to reject H_0, i.e., the value of 42.72 is not significantly greater than the established mean of 41.6 for non-users.*

t-test assumptions

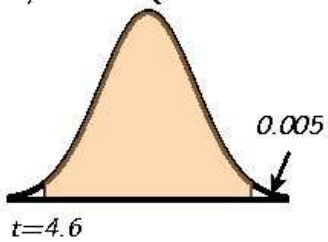
- Random sampling (like in z-test)
- Normal population (unlike z-test, where sample mean is automatically normal regardless of the population when sample size is large)
- Degrees of freedom = number of independent observations (actually, residuals)

Problem 6.2

A hospital exercise laboratory technician notes the resting pulse rates of five joggers to be 60, 58, 59, 61, and 67, respectively, while the resting pulse rates of seven non-exercisers are 83, 60, 75, 71, 91, 82, and 84, respectively. Establish a 99% confidence interval estimate for the difference in pulse rates between joggers and non-exercisers (means and standard deviations are: 61, 78, 3.54, and 10.23, respectively).

Solution

- $\mu_1 - \mu_2 = \overline{X} - \overline{Y} \pm t_{\alpha/2}(df)\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$
- $\alpha/2 = (1 - 0.99)/2 = 0.005$, $df = \min\{n_1, n_2\} - 1 = 5 - 1 = 4$



- $\mu_1 - \mu_2 = 17 \pm 4.6\sqrt{\frac{3.54^2}{5} + \frac{10.23^2}{7}} = 17 \pm 19 = [-2; 36]$
- *We are 99% confident that the true difference in pulse rates is between -2 and 36 bpm.*

Equal Variances Assumption

Assume that both populations have the same standard deviation (i.e., amount of exercise affects mean of the population, not its standard deviation)

$$\text{SE}(\overline{X} - \overline{Y}) = \sqrt{\frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}} \qquad df = \min\{n, m\} - 1$$

$$\text{SE}(\overline{X} - \overline{Y}) = \sigma\sqrt{\frac{1}{n} + \frac{1}{m}}, \text{if } \sigma_X = \sigma_Y \qquad df = n + m - 2$$

$$\hat{\sigma} = s_p = \sqrt{\frac{(n-1)s_X^2 + (m-1)s_Y^2}{n + m - 2}}$$

Problem 7.1

A hospital exercise laboratory technician notes the resting pulse rates of five joggers to be 60, 58, 59, 61, and 67, respectively, while the resting pulse rates of seven non-exercisers are 83, 60, 75, 71, 91, 82, and 84, respectively. Establish a 99% confidence interval estimate for the difference in pulse rates between joggers and non-exercisers (means and standard deviations are: 61, 78, 3.54, and 10.23, respectively). Assume equal variances.

Solution

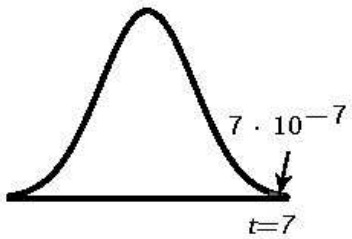
- $\mu_1 - \mu_2 = \overline{X} - \overline{Y} \pm t_{\alpha/2}(df)\sqrt{\frac{s_p^2}{n_1} + \frac{s_p^2}{n_2}}$
- $s_1 = s_2 = s_p = \sqrt{\frac{(n-1)s_X^2 + (m-1)s_Y^2}{n+m-2}} = \sqrt{\frac{4 \cdot 3.54^2 + 6 \cdot 10.23^2}{5+7-2}} = 8.23$
- $df = n_1 + n_2 - 2 = 5 + 7 - 2 = 10, \quad t_{0.005}(10) = 3.17$
- $\mu_1 - \mu_2 = 17 \pm 3.17 \cdot 8.23\sqrt{\frac{1}{5} + \frac{1}{7}} = 17 \pm 15 = [2; 32]$
- *We are 99% confident that the pulse rate difference is between 2 and 32 bpm.*

Problem 7.2

A researcher believes a new diet should improve weight gain in laboratory mice. If ten control mice on the old diet gain an average of 4 ounces with a standard deviation of 0.3 ounces, while the average gain for the ten mice on the new diet is 4.8 ounces with a standard deviation of 0.2 ounces, what is the P-value*?*

Solution

- $H_0 : \mu_2 = \mu_1$
- $H_a : \mu_2 > \mu_1$
- $s_p = \sqrt{\frac{0.3^2}{10} + \frac{0.2^2}{10}} = 0.255$, $df = 10 + 10 - 2 = 18$
- P-value $= \mathrm{P}(\overline{Y} - \overline{X} > 4.8 - 4) = \mathrm{P}(t(18) > \frac{0.8-0}{0.255\sqrt{\frac{1}{10}+\frac{1}{10}}}) = \mathrm{P}(t > 7) = 7 \cdot 10^{-7}$



- *At any reasonable significance level H_0 is rejected, i.e., there is strong evidence that new diet provides larger weight gain than the old diet.*

Dependent samples

Problem 7.3

Trace metals in drinking water wells affect the flavor of the water and unusually high concentrations can pose a health hazard. In the paper, "Trace Metals of South Indian River Region" (Environmental Studies, 1982, 62-6), trace metal concentrations (mg/L) on zinc were found from water drawn from the bottom and the top of each of 6 wells

Location	*Bottom*	*Top*
1	*0.430*	*0.415*
2	*0.266*	*0.238*
3	*0.567*	*0.390*
4	*0.531*	*0.410*
5	*0.707*	*0.605*
6	*0.716*	*0.609*

Solution

The two samples are dependent by construction.

Dependent samples

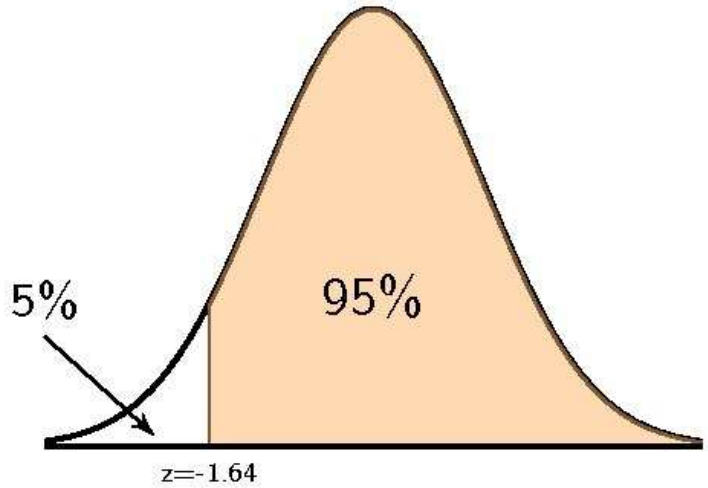
Location	Bottom	Top	Bottom - Top
1	0.430	0.415	0.015
2	0.266	0.238	0.028
3	0.567	0.390	0.177
4	0.531	0.410	0.121
5	0.707	0.605	0.102
6	0.716	0.609	0.107
		Mean	0.0916667
		SD	0.0606883

One-sample t-test:

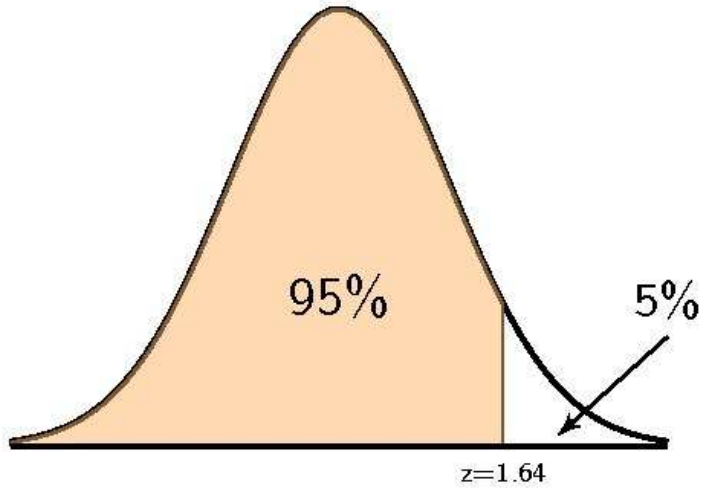
$\mathrm{P}(\overline{X} > 0.0917) = \mathrm{P}(t(5) > \frac{0.0917-0}{0.0607}) = \mathrm{P}(t(5) > 1.51) = 0.09 > 0.05$

At 5% significance level H_0 is not rejected, i.e., there is not enough evidence that more zinc is found on the bottom than on the top.

One-sided confidence intervals



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

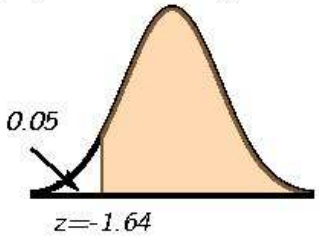


$$\mu < \overline{X} + z_\alpha \frac{\sigma}{\sqrt{n}}$$

Problem 8.1 (problem 3.1 revisited again)

A patient claims that he consumes only 2000 calories per day, but a dietician suspects that the actual figure is higher. The dietician checked his food intake for 30 days and found that the 30-day-mean is more than 2100 calories. Construct a **one-sided** *95% confidence interval for the number of calories in patient's diet. Assume standard deviation of 350 calories per day.*

Solution

- $\mu > \overline{X} - z_\alpha \cdot \text{SE}$, *where* $\alpha = 0.05$
- 

- $\mu > 2100 - 1.64 \cdot \frac{350}{\sqrt{30}} = [1995, +\infty)$
- *We are 95% confident that the value of μ is greater than 1995 cal*

Finite population size correction factor

- The formula $\sigma(\overline{X}) = \frac{\sigma}{\sqrt{n}}$ assumes that $X_1, \ldots, X_n$ is a sample of *independent* observations
- That is, $X_1, \ldots, X_n$ is a sample with replacement
- There is no difference between sampling with and without replacement if $n << N$, where N is the population size
- If population size and sample size are comparable, a correction factor is needed for $\sigma(\overline{X})$

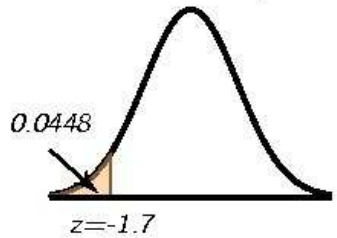
$$\sigma(\overline{X}) = \frac{\sigma}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}}$$

Problem 9.1 (Problem 3.3 revisited)

A coffee machine is filled with 200 coffee capsules and calibrated to deliver 8 ounces of coffee per cup. A random sample of 50 cups has the mean of 7.75 ounces and standard deviation of 1.2 ounces. Is there a reason to believe that the coffee machine was not calibrated well?

Solution

- $H_0 : \mu = 8$
- $H_a : \mu < 8$
- $\sigma(\overline{X}) = \frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}} = \frac{1.2}{\sqrt{50}}\sqrt{\frac{200-50}{200-1}} = 0.1473$
- P-value $= \mathrm{P}(\overline{X} < 7.75) = \mathrm{P}(Z < \frac{7.75-8}{0.1471}) = \mathrm{P}(Z < -1.7) = 0.0448$



- P-value $= 0.0448 < \alpha = 0.05$, *i.e., there is sufficient evidence at the 5% significance level to claim that the machine was not calibrated well.*

FAQs

- Do I have to divide by square root of n?
  - Yes, if I am computing $P(\overline{X} > 100)$
  - No, if I am computing $P(X > 100)$
- Do I have to divide by square root of n in one-proportion or two-proportion tests?
  - No. If you use Standard Error, it already contains the square root of n.
- When I compute standard deviation from the sample, do I have to divide it by square root of n?
  - Yes, if your calculations involve sample mean.

Remember that

- Sample standard deviation is an estimator for the population standard deviation
- Standard deviation of the sampling distribution is smaller than the population standard deviation
- Sample standard deviation is NOT an estimator for the standard deviation of the sampling distribution

Summary

- P-value is the probability of obtaining a result at least as extreme as the one that was actually observed, given that the null hypothesis is true
- Probability that the null hypothesis is true makes no sense
- An estimator is a statistic that is used to estimate an unknown population parameter
- Standard error is the standard deviation of the estimator
- Confidence intervals are random and depend on the sample
- t-test is used for small samples when and population variance is unknown
- Equal variances assumption increases effective sample size and gives less conservative estimates
- Finite population size correction factor accounts for sampling without replacement