Applied Biostatistics for Life Sciences

Dmitri D. Pervouchine

Skolkovo Institute for Science and Technology

Part 2: Statistical Inference

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

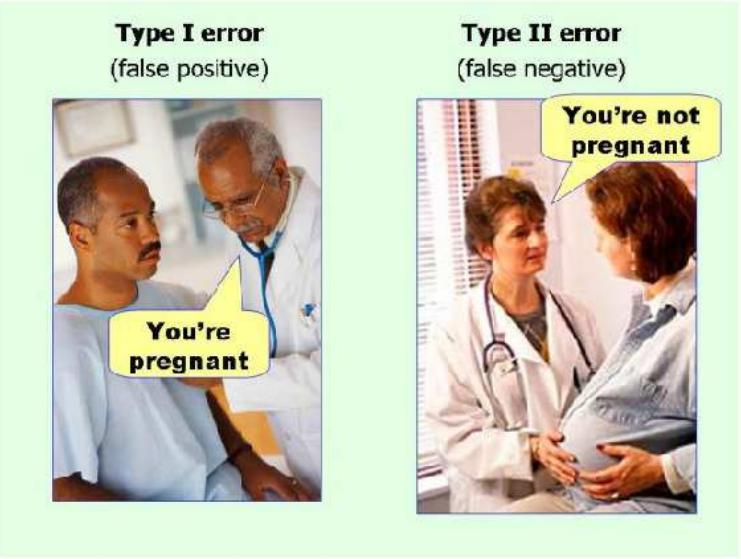
We believe or not believe in it!

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

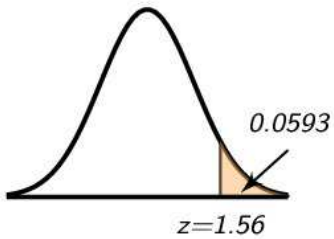
Meriam Webster Dictionary

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* $\overline{X} > 2100$
- $\mathrm{P}(\overline{X} > 2100) = \mathrm{P}(Z > \frac{2100-2000}{350/\sqrt{30}}) = \mathrm{P}(Z > 1.56) = 0.0593$



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

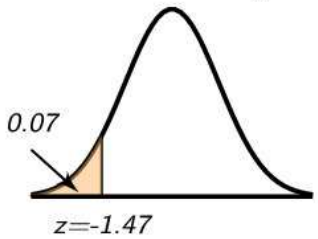
- P-value is a function of a sample
- α is a function of a decision rule
- Reject H_0 if P-value $< \alpha$
- Small P-value indicates that what you see would have been very unusual if H_0 were true

Problem 3.2

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 $= \mathrm{P}(\overline{X} < 7.75) = \mathrm{P}(Z < \frac{7.75-8}{0.1697}) = \mathrm{P}(Z < -1.47) = 0.07$



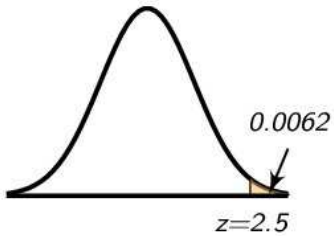
- P-value $= 0.07 > \alpha = 0.05$, *i.e., there is no sufficient evidence against the 8 ounces claim. That is, it is not too surprising to obtain a sample with* $\overline{X} = 7.75$ *if the actual* μ *were 8 oz.*

Problem 3.3

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

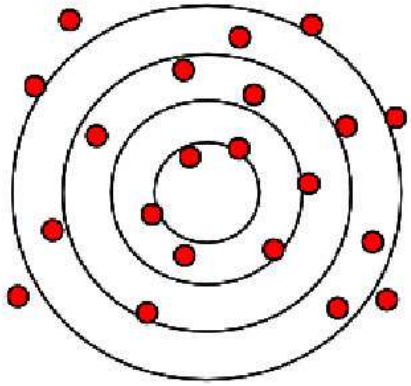
An *estimator* is a function of the observable sample data that is used to estimate an unknown population parameter

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

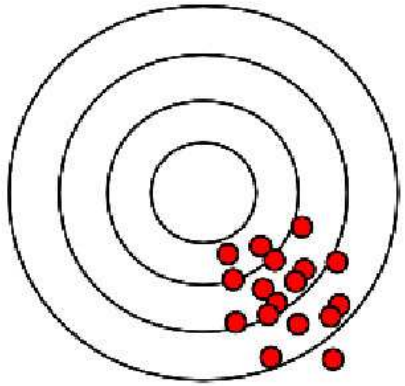
Unbiased and Effective Estimators

- 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

Standard Error

Standard error is the standard deviation of the estimator

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

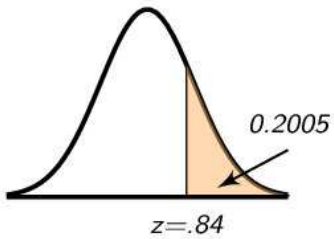
$$\mathrm{SE}(\hat{p}) = \sqrt{\frac{p(1-p)}{n}}$$

Problem 3.4

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* $\hat{p} > 0.18$
- P-value $= \mathrm{P}(\hat{p} > 0.18) = \mathrm{P}(Z > \frac{0.18-0.15}{\sqrt{\frac{0.15 \cdot 0.85}{100}}}) = \mathrm{P}(Z > 0.84) = 0.2005$



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

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

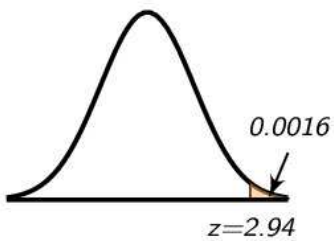
$$\mathrm{SE}(\hat{p}_1 - \hat{p}_2) = \sqrt{p(1-p)}\sqrt{\frac{1}{n} + \frac{1}{m}}, \text{ if } p_1 = p_2.$$

Problem 3.5

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

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

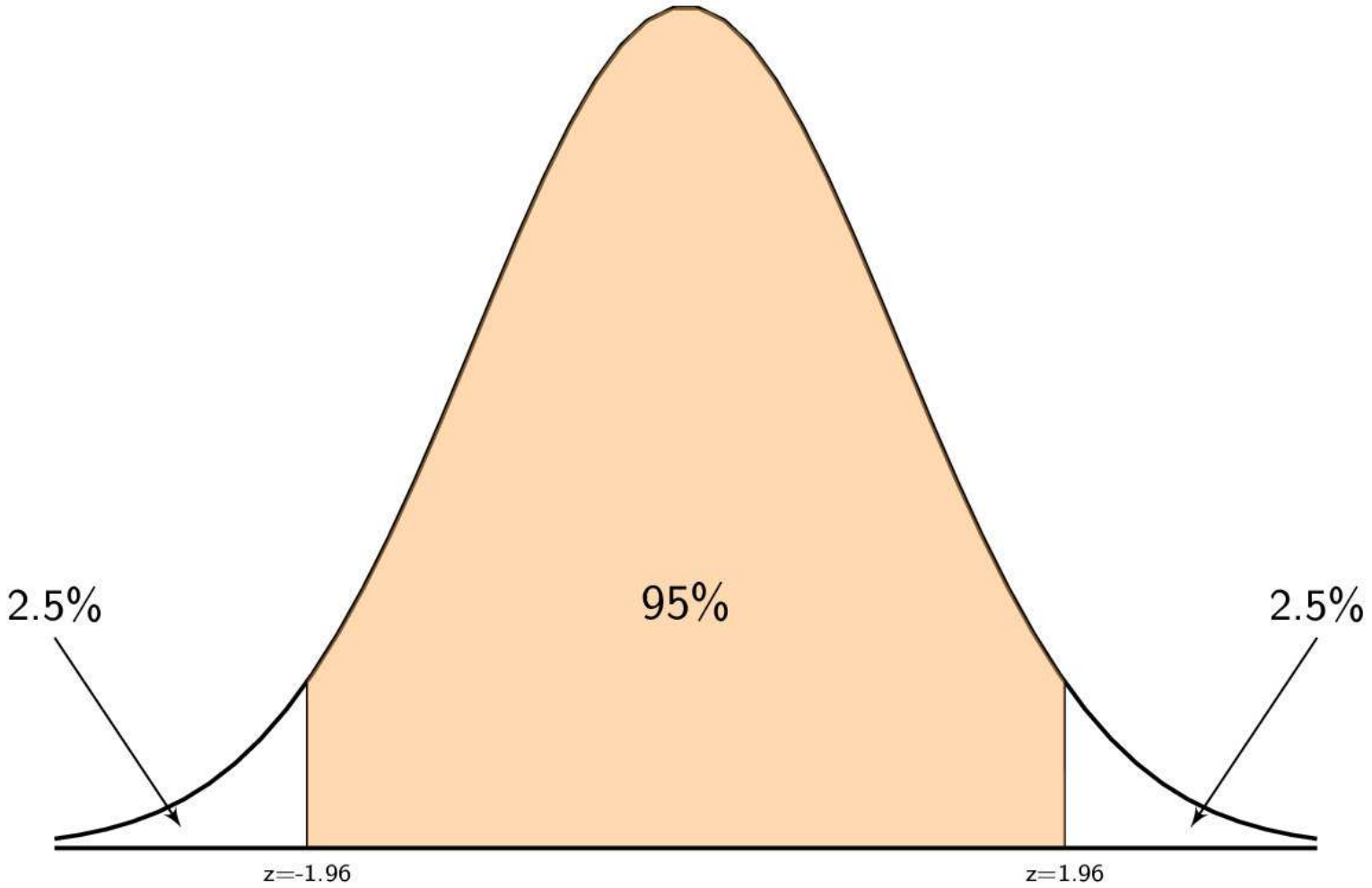
$$\text{SE} = \text{Standard Error}$$

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

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

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

Critical value

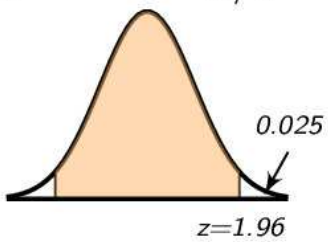


Problem 4.1 (problem 3.1 revisited)

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

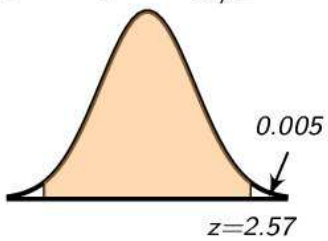
- We are 95% confident that the value of μ is in the confidence interval that we built
- 95% of samples result in confidence intervals which contain the true value of μ
- If we believe that our sample is "typical", i.e., within those 95% of samples, then yes the confidence interval that we built contains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

Problem 4.2 (problem 3.4 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* $\alpha/2 = (1 - 0.99)/2 = 0.005$



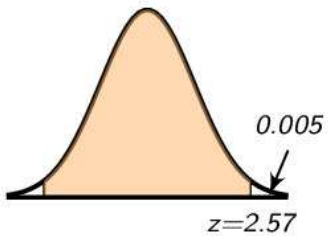
- $p = 0.18 \pm 2.57 \sqrt{\frac{0.18 \cdot 0.82}{100}} = 0.18 \pm 0.09 = [9\%, 27\%]$
- *We are 99% confident that between 9% and 27% of visitors stay for more than 2 days.*
- *Note that we replace p by $\hat{p}$ for the purpose of computing standard error.*

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

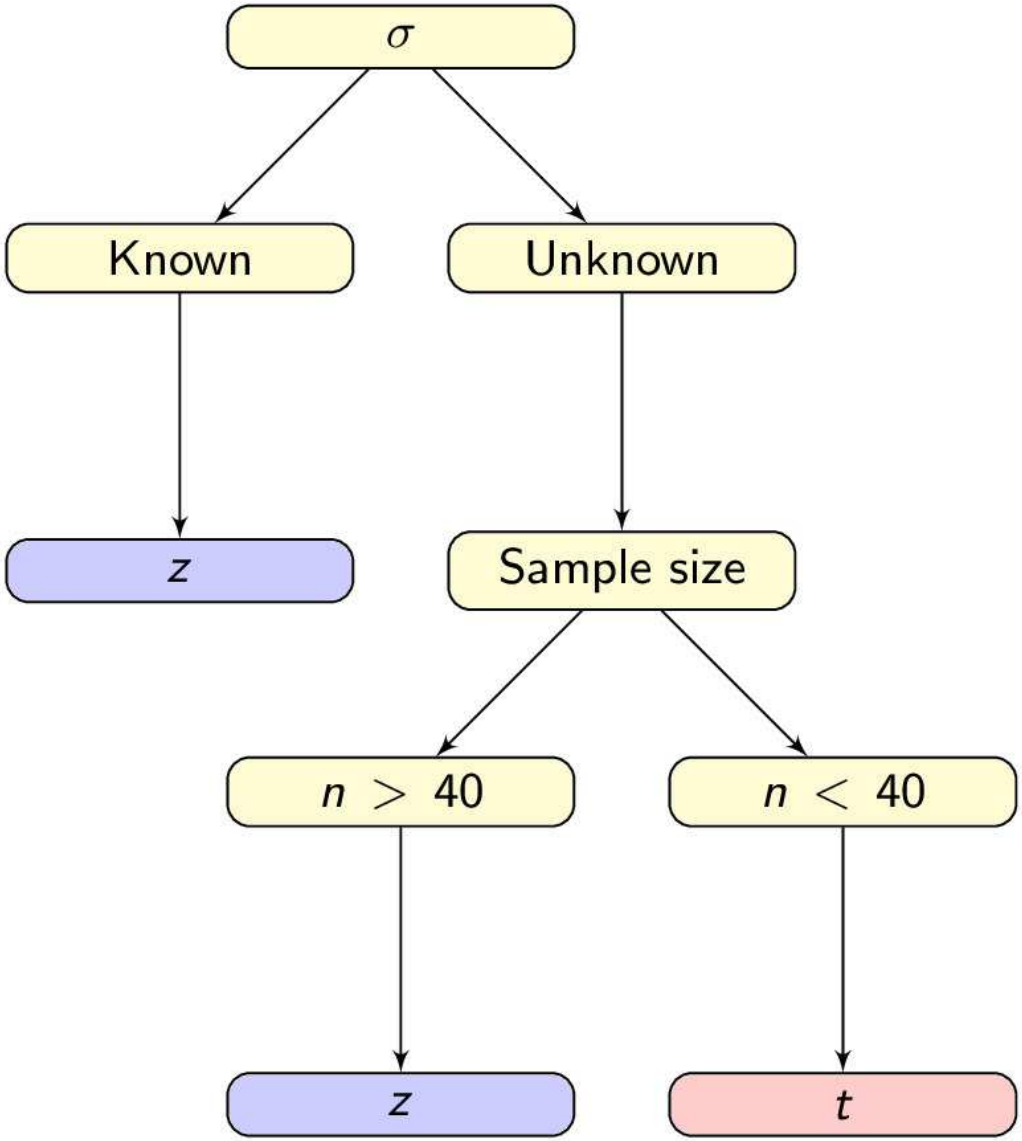


- $p_1 - p_2 = (0.75 - 0.4) \pm 2.57 \cdot 0.0389 = 0.35 \pm 0.10 = [25\%, 45\%]$
- *We are 99% confident that the proportion difference is between 25% and 45%.*

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



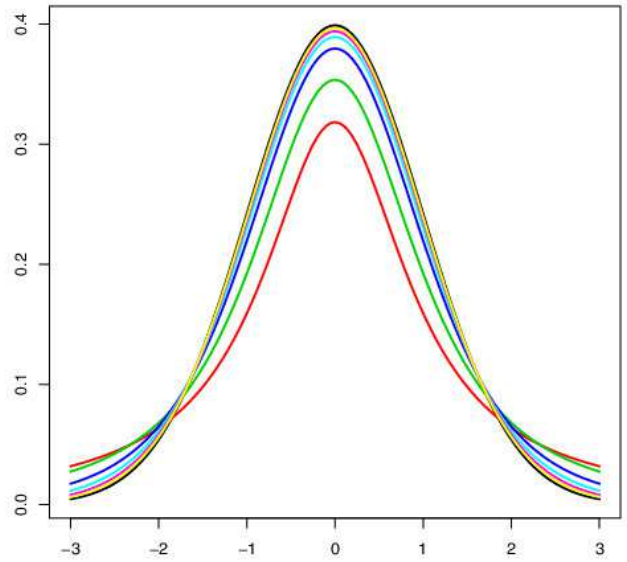
Student t-distribution

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

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

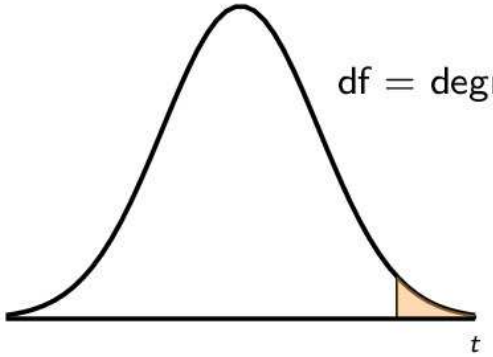
Student t-distribution

- 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

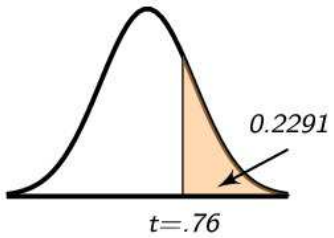
	Tail probability						
df	0.05	0.025	0.01	0.005	0.0025	0.001	0.0005
1	6.3138	12.7065	31.8193	63.6551	127.3447	318.4930	636.0450
2	2.9200	4.3026	6.9646	9.9247	14.0887	22.3276	31.5989
3	2.3534	3.1824	4.5407	5.8408	7.4534	10.2145	12.9242
4	2.1319	2.7764	3.7470	4.6041	5.5976	7.1732	8.6103
5	2.0150	2.5706	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

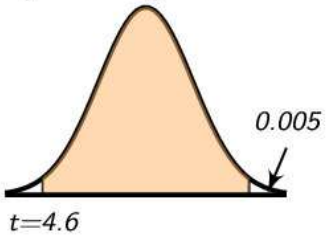
- Random sampling (like in z-test)
- Normal population (unlike z-test, where sample mean is automatically normal regardless of the population when sample size is large)
- Degrees of freedom = number of independent observations (actually, residuals)

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

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

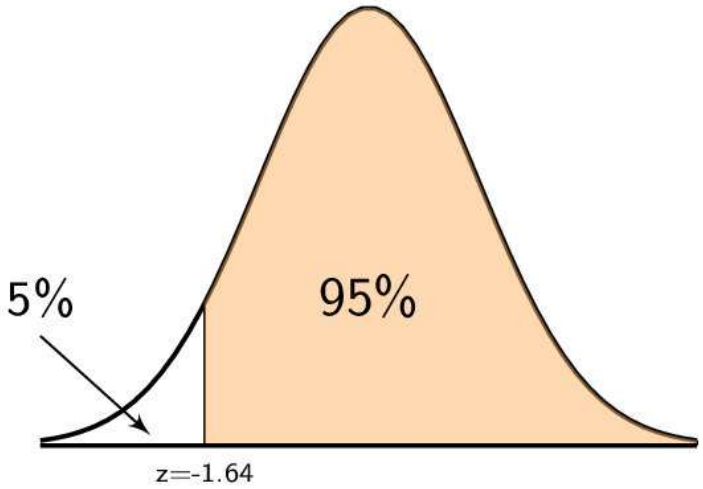
Location	Bottom	Top	Bottom - Top
1	0.430	0.415	0.015
2	0.266	0.238	0.028
3	0.567	0.390	0.177
4	0.531	0.410	0.121
5	0.707	0.605	0.102
6	0.716	0.609	0.107
		Mean	0.0916667
		SD	0.0606883

One-sample t-test:

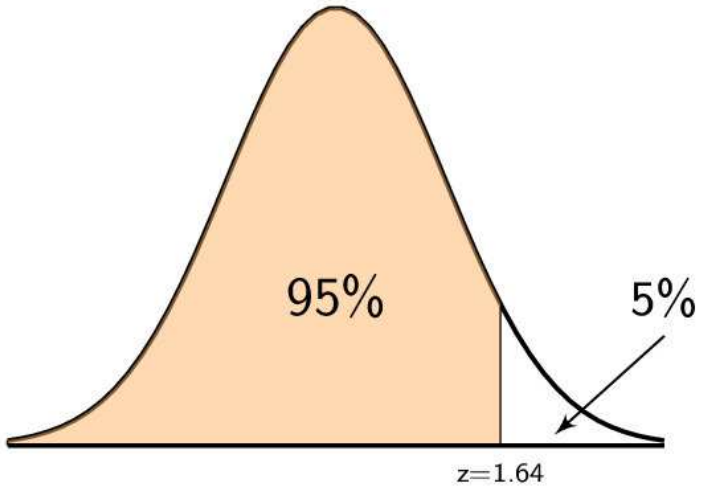
$$\mathrm{P}(\overline{X} > 0.0917) = \mathrm{P}(t(5) > \tfrac{0.0917-0}{0.0607/\sqrt{6}}) = \mathrm{P}(t(5) > 3.7) = 0.007 < 0.05$$

At 5% significance level H_0 is rejected, i.e., there is enough evidence that more zinc is found on the bottom than on the top.

One-sided confidence intervals



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



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

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 $\sigma(\overline{X})$

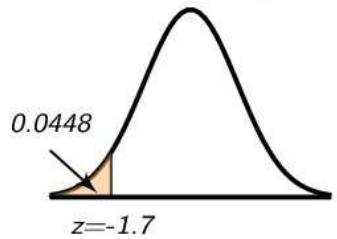
$$\sigma(\overline{X}) = \frac{\sigma}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}}$$

Problem 9.1 (Problem 3.2 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.*

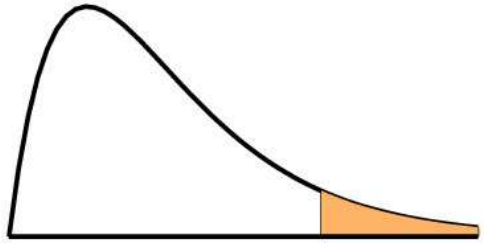
Estimation of Variance

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

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

$$\chi^2_{n-1} = \frac{s^2(n-1)}{\sigma^2}$$

Chi-square table



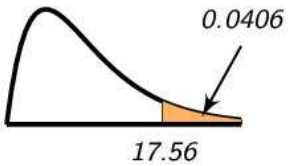
	Tail Probability $P(\chi^2 \geq a)$									
df	0.995	0.99	0.975	0.95	0.90	0.10	0.05	0.025	0.01	0.005
1	—	—	0.001	0.004	0.016	2.706	3.841	5.024	6.635	7.879
2	0.010	0.020	0.051	0.103	0.211	4.605	5.991	7.378	9.210	10.597
3	0.072	0.115	0.216	0.352	0.584	6.251	7.815	9.348	11.345	12.838
4	0.207	0.297	0.484	0.711	1.064	7.779	9.488	11.143	13.277	14.860
5	0.412	0.554	0.831	1.145	1.610	9.236	11.070	12.833	15.086	16.750
6	0.676	0.872	1.237	1.635	2.204	10.645	12.592	14.449	16.812	18.548
7	0.989	1.239	1.690	2.167	2.833	12.017	14.067	16.013	18.475	20.278
8	1.344	1.646	2.180	2.733	3.490	13.362	15.507	17.535	20.090	21.955
9	1.735	2.088	2.700	3.325	4.168	14.684	16.919	19.023	21.666	23.589
10	2.156	2.558	3.247	3.940	4.865	15.987	18.307	20.483	23.209	25.188
11	2.603	3.053	3.816	4.575	5.578	17.275	19.675	21.920	24.725	26.757
12	3.074	3.571	4.404	5.226	6.304	18.549	21.026	23.337	26.217	28.300
13	3.565	4.107	5.009	5.892	7.042	19.812	22.362	24.736	27.688	29.819
14	4.075	4.660	5.629	6.571	7.790	21.064	23.685	26.119	29.141	31.319
15	4.601	5.229	6.262	7.261	8.547	22.307	24.996	27.488	30.578	32.801

Problem 10.1

A supplier of 100 ohm/cm silicon wafers claims that his fabrication process can produce wafers with sufficient consistency so that the standard deviation of resistance for the lot does not exceed 10 ohm/cm. A sample of 10 wafers taken from the lot has a standard deviation of 13.97 ohm/cm. Is the suppliers claim reasonable?

Solution

- $H_0 : \sigma = 10$
- $H_a : \sigma > 10$
- $df = 10 - 1 = 9, \quad \mathrm{P}(s^2 > 13.97^2) = \mathrm{P}(\chi^2(9) > \frac{9 * 13.97^2}{10^2}) = \mathrm{P}(\chi^2(9) > 17.56) = 0.0406$



- *At 5% significance level the suppliers claim doesn't seem reasonable, i.e., there is enough reason to believe that* $\sigma > 10$.

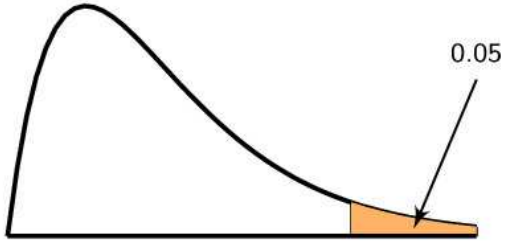
Fisher F-distribution

- $\frac{s_1^2(n_1-1)}{\sigma_1^2} \sim \chi^2_{n_1-1}$
- $\frac{s_2^2(n_2-1)}{\sigma_2^2} \sim \chi^2_{n_2-1}$
- $\frac{s_1^2/\sigma_1^2}{s_2^2/\sigma_2^2} \sim \frac{\frac{1}{n_1-1}\chi^2_{n_1-1}}{\frac{1}{n_2-1}\chi^2_{n_2-1}} = F(n_1-1, n_2-1)$
- The F-distribution is the ratio of two independent χ^2 variables divided by their respective degrees of freedom
- The F-test is designed to test if two population variances are equal

$$H_0 : \sigma_1^2 = \sigma_2^2$$

$$H_a : \sigma_1^2 \neq \sigma_2^2$$

Fisher F-distribution



$F(df_1, df_2)$

df_1	$df_2 = 2$	2	3	4	5	6	7	8	9	10
1	161.45	18.51	10.13	7.71	6.61	5.99	5.59	5.32	5.12	4.96
2	199.50	19.00	9.55	6.94	5.79	5.14	4.74	4.46	4.26	4.10
3	215.71	19.16	9.28	6.59	5.41	4.76	4.35	4.07	3.86	3.71
4	224.58	19.25	9.12	6.39	5.19	4.53	4.12	3.84	3.63	3.48
5	230.16	19.30	9.01	6.26	5.05	4.39	3.97	3.69	3.48	3.33
6	233.99	19.33	8.94	6.16	4.95	4.28	3.87	3.58	3.37	3.22
7	236.77	19.35	8.89	6.09	4.88	4.21	3.79	3.50	3.29	3.14
8	238.88	19.37	8.85	6.04	4.82	4.15	3.73	3.44	3.23	3.07
9	240.54	19.38	8.81	6.00	4.77	4.10	3.68	3.39	3.18	3.02
10	241.88	19.40	8.79	5.96	4.74	4.06	3.64	3.35	3.14	2.98

$$\mathrm{P}\left(F(df_1, df_2) < x\right) = \mathrm{P}\left(\frac{1}{F(df_1, df_2)} > \frac{1}{x}\right) = \mathrm{P}\left(F(df_2, df_1) > \frac{1}{x}\right)$$

Problem 11.1 (Exercise laboratory problem revisited)

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. The means and standard deviations for these samples are 61, 78, 3.54, and 10.23, respectively. Is equal variances assumption reasonable in this case?

Solution

- $H_0 : \sigma_1^2 = \sigma_2^2$
- $H_a : \sigma_1^2 \neq \sigma_2^2$
- $F = \frac{s_1^2/\sigma_1^2}{s_2^2/\sigma_2^2} = \frac{s_1^2}{s_2^2} = \frac{3.54^2}{10.23^2} = 0.12$
- $df_1 = 5 - 1 = 4;\ df_2 = 7 - 1 = 6$
- $\mathrm{P}(F(4,6) < 0.12) = \mathrm{P}(F(6,4) > \frac{1}{0.12}) = \mathrm{P}(F(6,4) > 8.35) < 0.05$ *since* $F_{0.05}(6,4) = 6.16$
- *There is enough evidence to reject* H_0 *at the 5% significance level, i.e., equal variances assumption is unreasonable.*

Estimation of Sample Size

- What is a minimum sample size needed to estimate the population mean within 2 units?

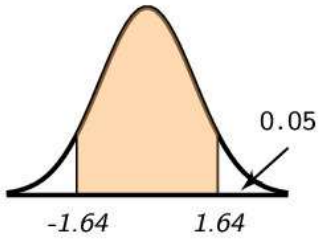
- What is a minimum sample size needed to estimate the population proportion within 2 percent units?

Problem 12.1

An electrical firm which manufactures a certain type of bulb wants to estimate its mean life. Assuming that the life of the light bulb is normally distributed and that the standard deviation is known to be 40 hours, how many bulbs should be tested so that we can be 90% confident that the estimate of the mean will not differ from the true mean life by more than 10 hours?

Solution

- $\mu = \overline{X} \pm z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}$, *where* $z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}} = 10$



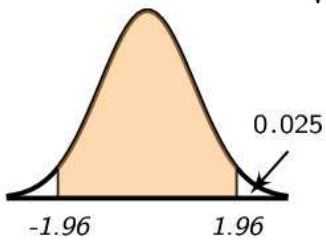
- $1.64 \frac{40}{\sqrt{n}} = 10$
- $n = 43.03 \rightarrow 44$

Problem 12.2

A quality control engineer wants to estimate the fraction of defective bulbs in a large lot of light bulbs. From past experience, he feels that the actual fraction of defective bulbs should be somewhere around 0.2 . How large a sample should be taken if he wants to estimate the true fraction within .02 using a 95% confidence interval?

Solution

- $p = \hat{p} \pm z_{\alpha/2} \cdot \sqrt{\frac{p(1-p)}{n}}$, *where* $z_{\alpha/2} \cdot \sqrt{\frac{p(1-p)}{n}} = 0.02$

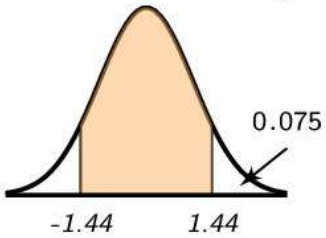


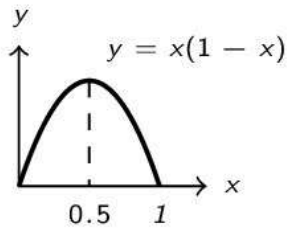
- $1.96\sqrt{\frac{0.2*0.8}{n}} = 0.02$
- $n = 1536.64 \rightarrow 1537$

Problem 12.3

Many television viewers express doubts about the validity of certain commercials. Let p represent the true proportion of consumers who believe what is shown in Timex television commercials. If Timex has no prior information regarding the true value of p, how many consumers should be included in their sample so that they will be 85% confident that their estimate is within 0.03 of the true value of p?

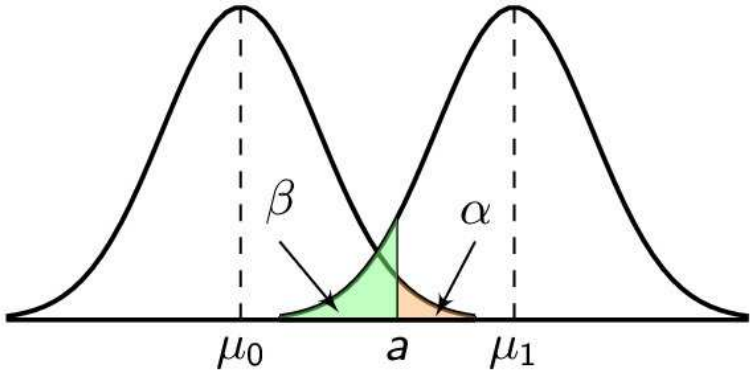
Solution

- $p = \hat{p} \pm z_{\alpha/2} \cdot \sqrt{\frac{p(1-p)}{n}}$, *where* $z_{\alpha/2} \cdot \sqrt{\frac{p(1-p)}{n}} = 0.03$
- 



$p = \frac{1}{2}$ *is the "worst" case*
- $1.44\sqrt{\frac{0.5*0.5}{n}} = 0.03$
- $n = 576$

Contribution of type I and type II errors



What is n such that the probability of committing type I error is α and the probability of committing type II error is β? The size of the effect is $\mu_1 - \mu_0 = \Delta$.

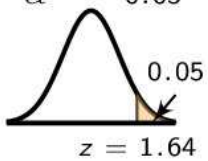
- $P(\overline{X} > a|\mu = \mu_0) = \alpha \quad P(\overline{X} < a|\mu = \mu_1) = \beta$
- $\begin{cases} \frac{a-\mu_0}{\sigma/\sqrt{n}} & = z_\alpha \\ \frac{\mu_1-a}{\sigma/\sqrt{n}} & = z_\beta \end{cases}$
- $a = \mu_0 + z_\alpha \frac{\sigma}{\sqrt{n}} = \mu_1 - z_\beta \frac{\sigma}{\sqrt{n}} \qquad (z_\alpha + z_\beta)\frac{\sigma}{\sqrt{n}} = \mu_1 - \mu_0 = \Delta$
- $n = \left(\frac{(z_\alpha + z_\beta)\sigma}{\Delta}\right)^2$

Problem 12.4

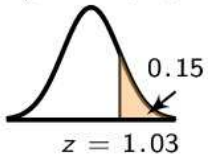
A clinical research organization is to design a pre-clinical of efficacy of a new drug to reduce the cholesterol level. The drug will be commercialized if the reduction of cholesterol be at least 2 mg/dL. Assuming the standard deviation of the cholesterol level in the target population is 20 mg/dL, what is the minimum sample size to achieve the desired reduction with at 5% significance level and with 15% type II error rate (85% power)?

Solution

- $z_\alpha = z_{0.05} = 1.64$



- $z_\beta = z_{0.15} = 1.03$



- $n = \left(\frac{(z_\alpha + z_\beta)\sigma}{\Delta}\right)^2 = \left(\frac{(1.64+1.03)20}{2}\right)^2 = 718.93 \to 719$

Taking into account finite population size

- $$\mu = \overline{X} \pm z\frac{\sigma}{\sqrt{n_0}}$$

  $$p = \hat{p} \pm z\sqrt{\frac{p(1-p)}{n_0}}$$

- $$\mu = \overline{X} \pm z\frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}}$$

  $$p = \hat{p} \pm z\sqrt{\frac{p(1-p)}{n}}\sqrt{\frac{N-n}{N-1}}$$

- $$n = \frac{n_0 N}{n_0 + N - 1}$$

Problem 12.5

An automobile dealer wants to estimate the proportion of customers who still own the cars they purchased 5 years earlier. Sales records indicate that the population of owners is 4,000. Set up a 95% confidence interval estimate of the population proportion of all customers who still own their cars 5 years after they were purchased if a random sample of 200 customers selected without replacement from the automobile dealer's records indicate that 82 still own cars that were purchased 5 years earlier. What sample size is necessary to estimate the true proportion to within ± 0.025 with 95% confidence?

Solution

- $p = \hat{p} \pm z\sqrt{\frac{p(1-p)}{n}}\sqrt{\frac{N-n}{N-1}}$
- $n = 200, N = 4000, \hat{p} = \frac{82}{200} = 0.41$
- $p = 0.41 \pm 0.07 * .97 \simeq 0.41 \pm 0.07$
- $n_0 = \frac{1.96^2 * 0.41 * 0.59}{0.025^2} = 1487$
- $n = \frac{n_0 N}{n_0 + N - 1} = \frac{1487 * 4000}{1487 + 4000 - 1} = 1085$

Chi-square Test for Independence

The test is applied when you have two categorical variables from a single population. It is used to determine whether there is a significant association between the two variables.

χ^2 test is applied to a contingency table with two factors

- H_0 : factors are independent
- H_a : factors are dependent

Problem 13.1

A restaurant owner surveys a random sample of 385 customers to determine whether customer satisfaction is related to gender and age.

	Young Male	*Young Female*	*Adult Male*	*Adult Female*
Satisfied	*25*	*30*	*135*	*112*
Not satisfied	*8*	*16*	*22*	*37*

Solution

	Young M	*Young F*	*Adult M*	*Adult F*	*Total*
Satisfied	*25*	*30*	*135*	*112*	*302*
Not satisfied	*8*	*16*	*22*	*37*	*83*
Total	*33*	*46*	*157*	*149*	*385*

If gender/age and satisfaction were independent then $\mathrm{P}(\textit{satisfied} \cap \textit{young male}) = \mathrm{P}(\textit{satisfied})\,\mathrm{P}(\textit{young male})$

Observed and Expected

- $P(\text{satisfied}) = 302/385$
- $P(\text{young male}) = 33/385$
- $P(\text{satisfied} \cap \text{young male}) = 302 * 33/385^2$
- Expected number of satisfied young males $= 302 * 33/385$

Observed:

	Young M	Young F	Adult M	Adult F	Total
Satisfied	25	30	135	112	302
Not satisfied	8	16	22	37	83
Total	33	46	157	149	385

Expected:

	Young M	Young F	Adult M	Adult F	Total
Satisfied	25.9	36.1	123.1	116.9	302
Not satisfied	7.1	9.9	33.9	32.1	83
Total	33	46	157	149	385

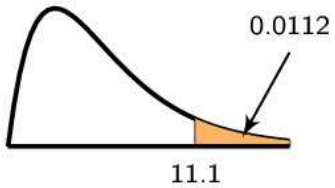
Chi-square Test for Independence

$$\chi^2 = \sum \frac{(O-E)^2}{E}$$

$$\chi^2 = \frac{(25-25.9)^2}{25.9} + \frac{(30-36.1)^2}{36.1} + \ldots = 11.1$$

$$df = (n-1)(m-1) = (2-1)(4-1) = 3$$

$$\mathrm{P}(\chi^2(3) \geq 11.1) = 0.112$$



At 5% significance level H_0 is rejected, i.e., there is evidence in this data that gender/age and satisfaction are not independent.

Chi-square Goodness of Fit

Problem 14.1

A grocery store manager wishes to determine whether a certain product will sell equally well in any of the five locations in the store. Five displays are set up, one for each location, and the resulting numbers of the product sold are noted

Location	*1*	*2*	*3*	*4*	*5*
Items sold	*43*	*29*	*52*	*34*	*48*

Is there enough evidence to claim a difference?

Solution

- H_0 : *The distribution is uniform*
- H_a : *The distribution is not uniform*
- *Total* $= 43+29+52+34+48=206$
- *We expect* $206/5=41.2$ *units sold in each location*

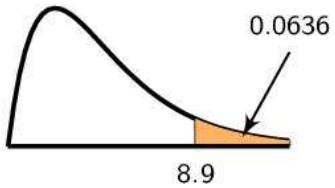
Chi-square Goodness of Fit

Location	1	2	3	4	5
Items sold	43	29	52	34	48
Expected	41.2	41.2	41.2	41.2	41.2

$$\chi^2 = \sum \frac{(O-E)^2}{E} = \frac{(43-41.2)^2}{41.2} + \ldots = 8.9$$

$$df = n - 1$$

$$\mathrm{P}(\chi^2(4) \geq 8.9) = 0.0636$$



At 5% significance level H_0 is not rejected, i.e., there is not enough evidence to claim that the five locations in the store are different.

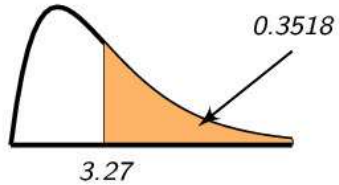
Problem 14.2

A geneticist claims that four species of fruit flies should appear in the ratio of 1:3:3:9. Suppose that a sample of 4000 fruit flies contained 226, 764, 733, and 2277 flies of each species, respectively. At the 10% significance level, is there sufficient evidence to reject the geneticist's hypothesis?

Solution

- $\frac{1}{16} + \frac{3}{16} + \frac{3}{16} + \frac{9}{16} = 1$, *that is* $4000 = 250 + 750 + 750 + 2250$

Observed	*226*	*764*	*733*	*2277*
Expected	*250*	*750*	*750*	*2250*



- $\chi^2 = \sum \frac{(O-E)^2}{E} = \frac{(226-250)^2}{250} + \frac{(764-750)^2}{750} + \ldots = 3.27$
- *The geneticist's hypothesis about 1:3:3:9 ratio is not rejected at any reasonable significance level, there is no reason to believe it is not true.*

Problem 14.3

Weights of rice bags are supposed to have normal distribution. A random sample of 40 such bags was taken and the following frequencies were obtained.

weight	*below 480*	*480-490*	*490-500*	*500-510*	*510-520*	*above 520*
number of bags	*6*	*9*	*10*	*8*	*4*	*3*

Test the hypothesis that rice bags were chosen from a normal distribution with the mean weight of 500 *grams and standard deviation of* 18 *grams.*

Solution

weight	*below 480*	*480-490*	*490-500*	*500-510*	*510-520*	*above 520*
	$z < -1.11$	$z \in (-1.11, -0.55]$	$z \in (-0.55, 0]$	$z \in (0, 0.55]$	$z \in (0.55, 1.11]$	$z > 1.11$
exp. prob	*0.1333*	*0.156*	*0.2107*	*0.2107*	*0.156*	*0.1333*
exp. count	*5.3*	*6.2*	*8.4*	*8.4*	*6.2*	*5.3*
observed	*6*	*9*	*10*	*8*	*4*	*3*

$$\chi^2 = \sum \frac{(O-E)^2}{E} = \frac{(6-5.3)^2}{5.3} + \cdots = 3.44$$

$\mathrm{P}(\chi^2(5) > 3.44) = 0.63$, *i.e., there is no evidence against the claim that rice bags were chosen from a normal distribution with the mean weight of* 500 *grams and standard deviation of* 18 *grams.*

Chi-square test: Warning

- Chi-square test is applicable only if the expected value in each cell is greater than 5 (Compare to Binomial Distribution)
- Small expected values lead to higher uncertainty in $\chi^2 = \sum \frac{(O-E)^2}{E}$
- You might find Fisher exact test (Hypergeometric test) also useful

Hypergeometric Test

Problem 14.4

A sample of teenagers might be divided into male and female on the one hand, and those that are and are not currently dieting on the other. We hypothesize, perhaps, that the proportion of dieting individuals is higher among the women than among the men, and we want to test whether any difference of proportions that we observe is significant.

	Men	*Women*	*Total*
Dieting	*1*	*9*	*10*
Not dieting	*11*	*3*	*14*
Total	*12*	*12*	*24*

Solution

	Men	*Women*	*Total*
Dieting	*5*	*5*	*10*
Not dieting	*7*	*7*	*14*
Total	*12*	*12*	*24*

Expected < 5

Hypergeometric Test

	Men	Women	Total
Dieting	1	9	10
Not dieting	11	3	14
Total	12	12	24

	Men	Women	Total
Dieting	a	b	$a+b$
Not dieting	c	d	$c+d$
Total	$a+c$	$b+d$	n

$$P = \frac{\binom{a+b}{a}\binom{c+d}{c}}{\binom{n}{a+c}} = \frac{(a+b)!(c+d)!(a+c)!(b+d)!}{n!a!b!c!d!}$$

$$P = \frac{10!14!12!12!}{24!1!9!11!3!} = 0.0013$$

Note that

- Exact computation with factorials of large numbers is troublesome
- Hypergeometric test is a **point** test, i.e., it estimates the probability of **exactly** the table that was observed. If you are interested in deviations in certain direction, you have to repeat hypergeometric test to compute hypergeometric CDF

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
- 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
- Finite population size correction factor accounts for sampling without replacement
- The ratio of sample variances of independent samples has Fisher F-distribution
- If population proportion is unknown, use $p = \frac{1}{2}$ to estimate sample size
- χ^2 test cannot be used when expected counts < 5