MATH 450: Mathematical statistics

November 12th, 2020

Lecture 20: P-values

Overview

Week 2	Chapter 6: Statistics and Sampling Distributions
Week 4	Chapter 7: Point Estimation
Week 7	Chapter 8: Confidence Intervals
Week 10	**Chapter 9: Tests of Hypotheses**
Week 12	Chapter 10: Two-sample testing
Week 14	Regression

Key steps in statistical inference

- Understand statistical models [Chapter 6]
- Come up with reasonable estimates of the parameters of interest [Chapter 7]
- Quantify the confidence with the estimates [Chapter 8]
- Testing with the parameters of interest [Chapter 9]

Contexts

- The central mega-example: population mean μ
- Difference between two population means

Chapter 9: Overview

9.1 Hypotheses and test procedures

- test procedures
- errors in hypothesis testing
- significance level

9.2 Tests about a population mean

- normal population with known σ
- large-sample tests
- a normal population with unknown σ

9.4 P-values

Hypothesis testing for one parameter

1. Identify the parameter of interest
2. Determine the null value and state the null hypothesis
3. State the appropriate alternative hypothesis
4. Give the formula for the test statistic
5. State the rejection region for the selected significance level α
6. Compute statistic value from data
7. Decide whether H_0 should be rejected and state this conclusion in the problem context

Test about a population mean

- Null hypothesis

$$H_0 : \mu = \mu_0$$

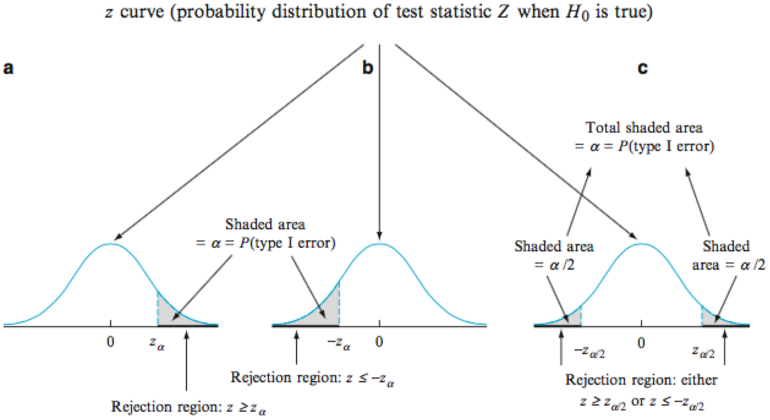
- The alternative hypothesis will be either:
    - $H_a : \mu > \mu_0$
    - $H_a : \mu < \mu_0$
    - $H_a : \mu \neq \mu_0$

Null hypothesis: $\mu = \mu_0$
Test statistic:

$$Z = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}}$$

Alternative Hypothesis	**Rejection Region for Level α Test**
H_a: $\mu > \mu_0$	$z \geq z_\alpha$ (upper-tailed test)
H_a: $\mu < \mu_0$	$z \leq -z_\alpha$ (lower-tailed test)
H_a: $\mu \neq \mu_0$	either $z \geq z_{\alpha/2}$ or $z \leq -z_{\alpha/2}$ (two-tailed test)

General rule



Example

Problem

A manufacturer of sprinkler systems used for fire protection in office buildings claims that the true average system-activation temperature is $130^\circ F$. *A sample of* $n = 9$ *systems, when tested, yields a sample average activation temperature of* $131.08^\circ F$.

If the distribution of activation times is normal with standard deviation $1.5^\circ F$, *does the data contradict the manufacturer's claim at significance level* $\alpha = 0.01$?

Solution

- Parameter of interest: $\mu =$ true average activation temperature
- Hypotheses

$$H_0 : \mu = 130$$
$$H_a : \mu \neq 130$$

- Test statistic:

$$z = \frac{\bar{x} - 130}{1.5/\sqrt{n}}$$

- Rejection region: either $z \leq -z_{0.005}$ or $z \geq z_{0.005} = 2.58$
- Substituting $\bar{x} = 131.08$, $n = 25 \rightarrow z = 2.16$.
- Note that $-2.58 < 2.16 < 2.58$. We fail to reject H_0 at significance level 0.01.
- The data does not give strong support to the claim that the true average differs from the design value.

Large-sample tests

Null hypothesis: $\mu = \mu_0$
Test statistic:

$$Z = \frac{\bar{X} - \mu_0}{S/\sqrt{n}}$$

Alternative Hypothesis	Rejection Region for Level α Test
H_a: $\mu > \mu_0$	$z \geq z_\alpha$ (upper-tailed test)
H_a: $\mu < \mu_0$	$z \leq -z_\alpha$ (lower-tailed test)
H_a: $\mu \neq \mu_0$	either $z \geq z_{\alpha/2}$ or $z \leq -z_{\alpha/2}$ (two-tailed test)

[Does not need the normal assumption]

Test about a normal population with unknown σ

Null hypothesis: H_0: $\mu = \mu_0$

Test statistic value: $t = \dfrac{\bar{x} - \mu_0}{s/\sqrt{n}}$

Alternative Hypothesis	**Rejection Region for a Level α Test**
H_a: $\mu > \mu_0$	$t \geq t_{\alpha,n-1}$ (upper-tailed)
H_a: $\mu < \mu_0$	$t \leq -t_{\alpha,n-1}$ (lower-tailed)
H_a: $\mu \neq \mu_0$	either $t \geq t_{\alpha/2,n-1}$ or $t \leq -t_{\alpha/2,n-1}$ (two-tailed)

[Require normal assumption]

Example

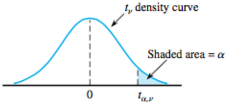
Problem

The amount of shaft wear (.0001 in.) after a fixed mileage was determined for each of $n = 8$ internal combustion engines having copper lead as a bearing material, resulting in $\bar{x} = 3.72$ and $s = 1.25$.

Assuming that the distribution of shaft wear is normal with mean μ, use the t-test at level 0.05 to test $H_0 : \mu = 3.5$ versus $H_a : \mu > 3.5$.

t-table

Table A.5 Critical Values for *t* Distributions



ν \ α	.10	.05	.025	.01	.005	.001	.0005
1	3.078	6.314	12.706	31.821	63.657	318.31	636.62
2	1.886	2.920	4.303	6.965	9.925	22.326	31.598
3	1.638	2.353	3.182	4.541	5.841	10.213	12.924
4	1.533	2.132	2.776	3.747	4.604	7.173	8.610
5	1.476	2.015	2.571	3.365	4.032	5.893	6.869
6	1.440	1.943	2.447	3.143	3.707	5.208	5.959
7	1.415	1.895	2.365	2.998	3.499	4.785	5.408
8	1.397	1.860	2.306	2.896	3.355	4.501	5.041
9	1.383	1.833	2.262	2.821	3.250	4.297	4.781
10	1.372	1.812	2.228	2.764	3.169	4.144	4.587
11	1.363	1.796	2.201	2.718	3.106	4.025	4.437
12	1.356	1.782	2.179	2.681	3.055	3.930	4.318
13	1.350	1.771	2.160	2.650	3.012	3.852	4.221
14	1.345	1.761	2.145	2.624	2.977	3.787	4.140
15	1.341	1.753	2.131	2.602	2.947	3.733	4.073
16	1.337	1.746	2.120	2.583	2.921	3.686	4.015
17	1.333	1.740	2.110	2.567	2.898	3.646	3.965
18	1.330	1.734	2.101	2.552	2.878	3.610	3.922
19	1.328	1.729	2.093	2.539	2.861	3.579	3.883
20	1.325	1.725	2.086	2.528	2.845	3.552	3.850
21	1.323	1.721	2.080	2.518	2.831	3.527	3.819
22	1.321	1.717	2.074	2.508	2.819	3.505	3.792
23	1.319	1.714	2.069	2.500	2.807	3.485	3.767
24	1.318	1.711	2.064	2.492	2.797	3.467	3.745

Practice

Problem

The standard thickness for silicon wafers used in a certain type of integrated circuit is 245 μm. A sample of 50 wafers is obtained and the thickness of each one is determined, resulting in a sample mean thickness of 246.18 μm and a sample standard deviation of 3.60 μm.
Does this data suggest that true average wafer thickness is larger than the target value? Carry out a test of significance at level .05.

Type II error and sample size determination

Hypothesis testing for one parameter

1. Identify the parameter of interest
2. Determine the null value and state the null hypothesis
3. State the appropriate alternative hypothesis
4. Give the formula for the test statistic
5. State the rejection region for the selected significance level α
6. Compute statistic value from data
7. Decide whether H_0 should be rejected and state this conclusion in the problem context

Type II error and sample size determination

- A level α test is a test with $P[\text{type I error}] = \alpha$
- Question: given α and n, can we compute β (the probabilities of type II error)?
- This is a very difficult question.
- We have a solution for the cases when: the distribution is normal **and** σ is known

Practice problem

Problem

The drying time of a certain type of paint under specified test conditions is known to be normally distributed with standard deviation 9 min. Assuming that we are testing

$$H_0 : \mu = 75$$
$$H_a : \mu < 75$$

from a dataset with $n = 25$.

- *What is the rejection region of the test with significance level $\alpha = 0.05$.*
- *What is $\beta(70)$ in this case?*

General cases

- Test of hypotheses:

$$H_0 : \mu = \mu_0$$
$$H_a : \mu < \mu_0$$

- Rejection region: $z \leq -z_\alpha$
- This is equivalent to $\bar{x} \leq \mu_0 - z_\alpha \sigma/\sqrt{n}$
- Let $\mu' < \mu_0$

$$\begin{aligned}
\beta(\mu') &= P[\text{Type II error when } \mu = \mu'] \\
&= P[H_0 \text{ is not rejected while it is false because } \mu = \mu'] \\
&= P[\bar{X} > \mu_0 - z_\alpha \sigma/\sqrt{n} \text{ while } \mu = \mu'] \\
&= P\left[\frac{\bar{X} - \mu'}{\sigma/\sqrt{n}} > \frac{\mu_0 - \mu'}{\sigma/\sqrt{n}} - z_\alpha \text{ while } \mu = \mu'\right] \\
&= 1 - \Phi\left(\frac{\mu_0 - \mu'}{\sigma/\sqrt{n}} - z_\alpha\right)
\end{aligned}$$

Remark

- For $\mu' < \mu_0$:

$$\beta(\mu') = 1 - \Phi\left(\frac{\mu_0 - \mu'}{\sigma/\sqrt{n}} - z_\alpha\right)$$

- If n, μ', μ_0, σ is fixed, then

$$\begin{aligned} &\beta(\mu') \text{ is small} \\ &\leftrightarrow \Phi\left(\frac{\mu_0 - \mu'}{\sigma/\sqrt{n}} - z_\alpha\right) \text{ is large} \\ &\leftrightarrow \frac{\mu_0 - \mu'}{\sigma/\sqrt{n}} - z_\alpha \text{ is large} \\ &\leftrightarrow \alpha \text{ is large} \end{aligned}$$

$\alpha - \beta$ compromise

Proposition

Suppose an experiment and a sample size are fixed and a test statistic is chosen. Then decreasing the size of the rejection region to obtain a smaller value of α *results in a larger value of* β *for any particular parameter value consistent with* H_a*.*

General formulas

Alternative Hypothesis	Type II Error Probability $\beta(\mu')$ for a Level α Test
H_a: $\mu > \mu_0$	$\Phi\left(z_\alpha + \frac{\mu_0 - \mu'}{\sigma/\sqrt{n}}\right)$
H_a: $\mu < \mu_0$	$1 - \Phi\left(-z_\alpha + \frac{\mu_0 - \mu'}{\sigma/\sqrt{n}}\right)$
H_a: $\mu \neq \mu_0$	$\Phi\left(z_{\alpha/2} + \frac{\mu_0 - \mu'}{\sigma/\sqrt{n}}\right) - \Phi\left(-z_{\alpha/2} + \frac{\mu_0 - \mu'}{\sigma/\sqrt{n}}\right)$

where $\Phi(z)$ = the standard normal cdf.

The sample size n for which a level α test also has $\beta(\mu') = \beta$ at the alternative value μ' is

$$n = \begin{cases} \left[\frac{\sigma(z_\alpha + z_\beta)}{\mu_0 - \mu'}\right]^2 & \text{for a one - tailed (upper or lower) test} \\ \left[\frac{\sigma(z_{\alpha/2} + z_\beta)}{\mu_0 - \mu'}\right]^2 & \text{for a two - tailed test (an approximate solution)} \end{cases}$$

P-values

Remarks

- The common approach in statistical testing is:
  1. specifying significance level α
  2. reject/not reject H_0 based on evidence
- Weaknesses of this approach:
  - it says nothing about whether the computed value of the test statistic just barely fell into the rejection region or whether it exceeded the critical value by a large amount
  - each individual may select their own significance level for their presentation
- We also want to include some *objective* quantity that describes how *strong* the rejection is $\rightarrow$ P-value

Practice problem

Problem

Suppose μ was the true average nicotine content of brand of cigarettes. We want to test:

$$H_0 : \mu = 1.5$$
$$H_a : \mu > 1.5$$

Suppose that $n = 64$ and $z = \frac{\bar{x}-1.5}{s/\sqrt{n}} = 2.1$. Will we reject H_0 if the significance level is

(a) $\alpha = 0.05$
(b) $\alpha = 0.025$
(c) $\alpha = 0.01$
(d) $\alpha = 0.005$

P-value

Level of Significance α	Rejection Region	Conclusion
.05	$z \geq 1.645$	Reject H_0
.025	$z \geq 1.96$	Reject H_0
.01	$z \geq 2.33$	Do not reject H_0
.005	$z \geq 2.58$	Do not reject H_0

Question: What is the smallest value of α for which H_0 is rejected.

P-value

DEFINITION The ***P*-value** (or *observed significance level*) is the smallest level of significance at which H_0 would be rejected when a specified test procedure is used on a given data set. Once the P-value has been determined, the conclusion at any particular level α results from comparing the P-value to α:

1. P-value $\leq \alpha \Rightarrow$ reject H_0 at level α.
2. P-value $> \alpha \Rightarrow$ do not reject H_0 at level α.

Testing by P-value method

DECISION RULE BASED ON THE *P*-VALUE	Select a significance level α (as before, the desired type I error probability). Then reject H_0 if P-value $\leq \alpha$; do not reject H_0 if P-value $> \alpha$

Remark: the smaller the P-value, the more evidence there is in the sample data against the null hypothesis and for the alternative hypothesis.

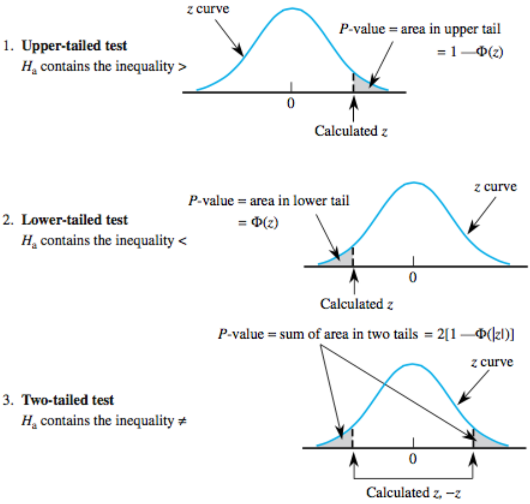


Figure 9.7 Determination of the P-value for a z test

Practice problem

Problem

The target thickness for silicon wafers used in a certain type of integrated circuit is 245 μm. A sample of 50 wafers is obtained and the thickness of each one is determined, resulting in a sample mean thickness of 246.18 μm and a sample standard deviation of 3.60 μm.
At confidence level $\alpha = 0.01$, does this data suggest that true average wafer thickness is something other than the target value?

$\Phi(z)$

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279	.5319	.5359
0.1	.5398	.5438	.5478	.5517	.5557	.5596	.5636	.5675	.5714	.5753
0.2	.5793	.5832	.5871	.5910	.5948	.5987	.6026	.6064	.6103	.6141
0.3	.6179	.6217	.6255	.6293	.6331	.6368	.6406	.6443	.6480	.6517
0.4	.6554	.6591	.6628	.6664	.6700	.6736	.6772	.6808	.6844	.6879
0.5	.6915	.6950	.6985	.7019	.7054	.7088	.7123	.7157	.7190	.7224
0.6	.7257	.7291	.7324	.7357	.7389	.7422	.7454	.7486	.7517	.7549
0.7	.7580	.7611	.7642	.7673	.7704	.7734	.7764	.7794	.7823	.7852
0.8	.7881	.7910	.7939	.7967	.7995	.8023	.8051	.8078	.8106	.8133
0.9	.8159	.8186	.8212	.8238	.8264	.8289	.8315	.8340	.8365	.8389
1.0	.8413	.8438	.8461	.8485	.8508	.8531	.8554	.8577	.8599	.8621
1.1	.8643	.8665	.8686	.8708	.8729	.8749	.8770	.8790	.8810	.8830
1.2	.8849	.8869	.8888	.8907	.8925	.8944	.8962	.8980	.8997	.9015
1.3	.9032	.9049	.9066	.9082	.9099	.9115	.9131	.9147	.9162	.9177
1.4	.9192	.9207	.9222	.9236	.9251	.9265	.9278	.9292	.9306	.9319
1.5	.9332	.9345	.9357	.9370	.9382	.9394	.9406	.9418	.9429	.9441
1.6	.9452	.9463	.9474	.9484	.9495	.9505	.9515	.9525	.9535	.9545
1.7	.9554	.9564	.9573	.9582	.9591	.9599	.9608	.9616	.9625	.9633
1.8	.9641	.9649	.9656	.9664	.9671	.9678	.9686	.9693	.9699	.9706
1.9	.9713	.9719	.9726	.9732	.9738	.9744	.9750	.9756	.9761	.9767
2.0	.9772	.9778	.9783	.9788	.9793	.9798	.9803	.9808	.9812	.9817
2.1	.9821	.9826	.9830	.9834	.9838	.9842	.9846	.9850	.9854	.9857
2.2	.9861	.9864	.9868	.9871	.9875	.9878	.9881	.9884	.9887	.9890
2.3	.9893	.9896	.9898	.9901	.9904	.9906	.9909	.9911	.9913	.9916
2.4	.9918	.9920	.9922	.9925	.9927	.9929	.9931	.9932	.9934	.9936
2.5	.9938	.9940	.9941	.9943	.9945	.9946	.9948	.9949	.9951	.9952
2.6	.9953	.9955	.9956	.9957	.9959	.9960	.9961	.9962	.9963	.9964
2.7	.9965	.9966	.9967	.9968	.9969	.9970	.9971	.9972	.9973	.9974
2.8	.9974	.9975	.9976	.9977	.9977	.9978	.9979	.9979	.9980	.9981
2.9	.9981	.9982	.9982	.9983	.9984	.9984	.9985	.9985	.9986	.9986
3.0	.9987	.9987	.9987	.9988	.9988	.9989	.9989	.9989	.9990	.9990
3.1	.9990	.9991	.9991	.9991	.9992	.9992	.9992	.9992	.9993	.9993
3.2	.9993	.9993	.9994	.9994	.9994	.9994	.9994	.9995	.9995	.9995
3.3	.9995	.9995	.9995	.9996	.9996	.9996	.9996	.9996	.9996	.9997

P-values for z-tests

1. Parameter of interest: μ = true average wafer thickness
2. Null hypothesis: H_0: $\mu = 245$
3. Alternative hypothesis: H_a: $\mu \neq 245$
4. Formula for test statistic value: $z = \dfrac{\bar{x} - 245}{s/\sqrt{n}}$
5. Calculation of test statistic value: $z = \dfrac{246.18 - 245}{3.60/\sqrt{50}} = 2.32$
6. Determination of P-value: Because the test is two-tailed,

$$P\text{-value} = 2[1 - \Phi(2.32)] = .0204$$

7. Conclusion: Using a significance level of .01, H_0 would not be rejected since $.0204 > .01$. At this significance level, there is insufficient evidence to conclude that true average thickness differs from the target value.

P-values for z-tests

$$P\text{-value:} \quad P = \begin{cases} 1 - \Phi(z) & \text{for an upper-tailed test} \\ \Phi(z) & \text{for a lower-tailed test} \\ 2[1 - \Phi(|z|)] & \text{for a two-tailed test} \end{cases}$$

P-values for t-tests

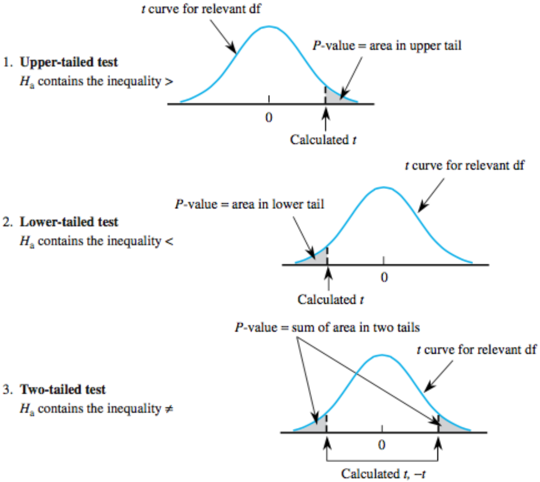


Figure 9.8 P-values for t tests

Practice problem

Problem

Suppose we want to test

$$H_0 : \mu = 25$$
$$H_a : \mu > 25$$

from a sample with $n = 5$ *and the calculated value*

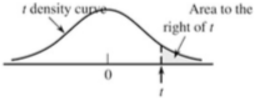
$$t = \frac{\bar{x} - 25}{s/\sqrt{n}} = 1.02$$

(a) *What is the P-value of the test*

(b) *Should we reject the null hypothesis?*

t-table

Table A.7 *t* Curve Tail Areas



t \ ν	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
0.0	.500	.500	.500	.500	.500	.500	.500	.500	.500	.500	.500	.500	.500	.500	.500	.500	.500	.500
0.1	.468	.465	.463	.463	.462.	.462	.462	.461	.461	.461	.461	.461	.461	.461	.461	.461	.461	.461
0.2	.437	.430	.427	.426	.425	.424	.424	.423	.423	.423	.423	.422	.422	.422	.422	.422	.422	.422
0.3	.407	.396	.392	.390	.388	.387	.386	.386	.386	.385	.385	.385	.384	.384	.384	.384	.384	.384
0.4	.379	.364	.358	.355	.353	.352	.351	.350	.349	.349	.348	.348	.348	.347	.347	.347	.347	.347
0.5	.352	.333	.326	.322	.319	.317	.316	.315	.315	.314	.313	.313	.313	.312	.312	.312	.312	.312
0.6	.328	.305	.295	.290	.287	.285	.284	.283	.282	.281	.280	.280	.279	.279	.279	.278	.278	.278
0.7	.306	.278	.267	.261	.258	.255	.253	.252	.251	.250	.249	.249	.248	.247	.247	.247	.247	.246
0.8	.285	.254	.241	.234	.230	.227	.225	.223	.222	.221	.220	.220	.219	.218	.218	.218	.217	.217
0.9	.267	.232	.217	.210	.205	.201	.199	.197	.196	.195	.194	.193	.192	.191	.191	.191	.190	.190
1.0	.250	.211	.196	.187	.182	.178	.175	.173	.172	.170	.169	.169	.168	.167	.167	.166	.166	.165
1.1	.235	.193	.176	.167	.162	.157	.154	.152	.150	.149	.147	.146	.146	.144	.144	.144	.143	.143
1.2	.221	.177	.158	.148	.142	.138	.135	.132	.130	.129	.128	.127	.126	.124	.124	.124	.123	.123
1.3	.209	.162	.142	.132	.125	.121	.117	.115	.113	.111	.110	.109	.108	.107	.107	.106	.105	.105
1.4	.197	.148	.128	.117	.110	.106	.102	.100	.098	.096	.095	.093	.092	.091	.091	.090	.090	.089
1.5	.187	.136	.115	.104	.097	.092	.089	.086	.084	.082	.081	.080	.079	.077	.077	.077	.076	.075

A P-value:

- is not the probability that H_0 is true
- is not the probability of rejecting H_0
- is the probability, calculated assuming that H_0 is true, of obtaining a test statistic value at least as contradictory to the null hypothesis as the value that actually resulted

Example 1

Let μ denote the mean reaction time to a certain stimulus. For a large-sample z test of H_0: $\mu = 5$ versus H_a: $\mu > 5$, nd the P-value associated with each of the given values of the z test statistic.

a. 1.42 **b.** .90 **c.** 1.96

d. 2.48 **e.** $-.11$

On the label, Pepperidge Farm bagels are said to weigh four ounces each (113 grams). A random sample of six bagels resulted in the following weights (in grams):

117.6 109.5 111.6 109.2 119.1 110.8

a. Based on this sample, is there any reason to doubt that the population mean is at least 113 grams?