Small-Sample Statistics

Analyzing normally distributed data when both μ and σ^2 are unknown

Bruce M. Boghosian

Tufts University School of Arts and Sciences

Department of Mathematics

Tufts University

Outline

1 Interval estimation

2 Hypothesis testing

3 Interval estimation of the variance

4 Hypothesis testing with the variance

5 Summary

Interval estimation of μ using the Z ratio

- We know that the Z ratio, $Z = \frac{\overline{Y}-\mu}{\sigma/\sqrt{n}}$ is distributed as a standard normal.
- Hence we can write

$$P\left(-z_{\alpha/2} \leq \frac{\overline{Y}-\mu}{\sigma/\sqrt{n}} \leq +z_{\alpha/2}\right) = 1-\alpha$$

$$\therefore \quad P\left(\overline{Y} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \leq \mu \leq \overline{Y} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right) = 1-\alpha.$$

- Hence, if $y_1, \ldots, y_n$ is random sample of size n from normal distribution with known σ and unknown μ, then a $100(1-\alpha)\%$ confidence interval for μ is

$$\left(\overline{y} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \overline{y} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right)$$

Interval estimation of μ using the T ratio

- We know that the T ratio, $T_{n-1} = \frac{\overline{Y}-\mu}{S/\sqrt{n}}$ is distributed as a Student T distribution with $n-1$ degrees of freedom.
- Hence we can write

$$P\left(-t_{\alpha/2,n-1} \leq \frac{\overline{Y}-\mu}{S/\sqrt{n}} \leq +t_{\alpha/2,n-1}\right) = 1-\alpha$$
$$\therefore \quad P\left(\overline{Y} - t_{\alpha/2,n-1}\frac{S}{\sqrt{n}} \leq \mu \leq \overline{Y} + t_{\alpha/2,n-1}\frac{S}{\sqrt{n}}\right) = 1-\alpha.$$

- Hence, if $y_1, \ldots, y_n$ is a random sample of size n from a normal distribution with unknown mean μ, then a $100(1-\alpha)\%$ confidence interval for μ is

$$\left(\overline{y} - t_{\alpha/2,n-1}\frac{s}{\sqrt{n}}, \overline{y} + t_{\alpha/2,n-1}\frac{s}{\sqrt{n}}\right)$$

Example

- Distance between bat and insect at the moment bat first detects insect (cm)

$$62, 52, 68, 23, 34, 45, 27, 42, 83, 56, 40$$

- Note $n = 11$, $\overline{y} = 48.36$, and $s = 18.08$
- Taking $\alpha = 0.05$, note that $t_{0.05/2,11-1} = t_{0.025,10} = 2.2281$
- Then the 95% confidence interval for μ is

$$\left(48.36 - 2.2281\frac{18.08}{\sqrt{11}}, 48.36 + 2.2281\frac{18.08}{\sqrt{11}}\right) = (36.21, 60.51)$$

Tufts UNIVERSITY

Hypothesis testing with the Z ratio

- Given data $Y_1, \ldots, Y_n$, drawn from a distribution that is known to be normal with known standard deviation σ_Y, various null hypotheses can be tested by using the fact that the Z ratio, $Z = \frac{\overline{Y} - \mu_0}{\sigma/\sqrt{n}}$ is normally distributed.
- So calculate $z = \frac{\overline{y} - \mu_0}{\sigma/\sqrt{n}}$, and...
  - To test $H_0 : \mu = \mu_0$ versus $H_1 : \ \mu > \mu_0$ at the α level of significance, reject H_0 if $z \geq +z_\alpha$.
  - To test $H_0 : \mu = \mu_0$ versus $H_1 : \ \mu < \mu_0$ at the α level of significance, reject H_0 if $z \leq -z_\alpha$.
  - To test $H_0 : \mu = \mu_0$ versus $H_1 : \ \mu \neq \mu_0$ at the α level of significance, reject H_0 if z is either (a) $\leq -z_{\alpha/2}$, or (b) $\geq +z_{\alpha/2}$.
- Problem with this approach: We usually do not have a priori knowledge of the standard deviation σ_Y.

Hypothesis testing with T ratio: One-sample t test

Bruce M. Boghosian

Interval estimation

Hypothesis testing

Interval estimation of the variance

Hypothesis testing with the variance

Summary

- Under the null hypothesis, we know that the T ratio, $T_{n-1} = \frac{\overline{Y}-\mu_0}{S/\sqrt{n}}$ is distributed as a Student T distribution with $n-1$ degrees of freedom.
- So calculate $t = \frac{\overline{y}-\mu_0}{s/\sqrt{n}}$, and...
  - To test $H_0 : \mu = \mu_0$ versus $H_1 : \ \mu > \mu_0$ at the α level of significance, reject H_0 if $t \geq +t_{\alpha,n-1}$.
  - To test $H_0 : \mu = \mu_0$ versus $H_1 : \ \mu < \mu_0$ at the α level of significance, reject H_0 if $t \leq -t_{\alpha,n-1}$.
  - To test $H_0 : \mu = \mu_0$ versus $H_1 : \ \mu \neq \mu_0$ at the α level of significance, reject H_0 if t is either (a) $\leq -t_{\alpha/2,n-1}$, or (b) $\geq +t_{\alpha/2,n-1}$.

Example

- Corrosion of metal pipe for underground cables in one year
- Without plastic coating, average pit depth is 0.0042 inch
- $n = 10$ tests with plastic coating yield numbers (in inches):

0.0039	0.0041	0.0038	0.0044	0.0040
0.0036	0.0034	0.0036	0.0046	0.0036

- Sample mean is $\mu_0 = 0.0039$ inch
- Sample standard deviation is $s = 0.000383$ inch
- Can we conclude, at the $\alpha = 0.05$ level of significance, that the plastic coating is beneficial?

Example (continued)

- H_0: Plastic coating has no effect, so $\mu = \mu_0 = 0.0042$ inch
- H_1: Plastic coating has beneficial effect, so $\mu < \mu_0 = 0.0042$ inch
- Calculate $t = \frac{\overline{y}-\mu_0}{s/\sqrt{n}} = \frac{0.0039-0.0042}{0.000383/\sqrt{10}} = -2.47717$
- We reject H_0 since $t < -t_{\alpha,n-1} = -t_{0.05,9} = -1.8331$.
- Conclude that plastic coating has some beneficial effect with 95% confidence.

An important distinction

- Note z tests can be used on data not normally distributed, as long as σ^2 is finite and n sufficiently large that CLT can be invoked to claim that $\overline{y}$ is normally distributed.
- To use t tests, one must be sure that each of the Y_j are normally distributed. The derivation of the Student T distribution depends on this assumption.
- Unfortunately, it is sometimes very difficult to know for sure the exact pdf of the data that you are measuring.
- Two heuristics for using the T test in such a situation
  - Histogram $\frac{\overline{Y}-\mu}{S/\sqrt{n}}$ to make sure that it is not too skewed.
  - When n is sufficiently large, the pdf of $\frac{\overline{Y}-\mu}{S/\sqrt{n}}$ becomes similar to that of $f_{T_{n-1}}(t)$.
  - t test is *robust with respect to departures from normality*, as is z test.

Constructing confidence intervals for σ^2

- How do we construct a confidence interval for σ^2?
- This is something that does not even come up for the z test, for which σ^2 is fixed and assumed known.
- We begin with two facts about the sample variance S^2
  - $S^2 = \frac{1}{n-1}\sum_{j=1}^{n}\left(Y_j - \overline{Y}\right)^2$ is an unbiased estimator for σ^2
  - $\frac{(n-1)S^2}{\sigma^2} = \frac{1}{\sigma^2}\sum_{j=1}^{n}\left(Y_j - \overline{Y}\right)^2$ is χ^2 with $n-1$ df
- It follows that

$$P\left[\chi^2_{\alpha/2,n-1} \le \frac{(n-1)S^2}{\sigma^2} \le \chi^2_{1-\alpha/2,n-1}\right] = 1-\alpha$$

$$\therefore \quad P\left[\frac{(n-1)s^2}{\chi^2_{1-\alpha/2,n-1}} \le \sigma^2 \le \frac{(n-1)s^2}{\chi^2_{\alpha/2,n-1}}\right] = 1-\alpha$$

Constructing confidence intervals for σ^2 (continued)

- The $100(1-\alpha)\%$ confidence interval for σ^2 is

$$\left(\frac{(n-1)s^2}{\chi^2_{1-\alpha/2,n-1}}, \frac{(n-1)s^2}{\chi^2_{\alpha/2,n-1}} \right)$$

- Likewise the $100(1-\alpha)\%$ confidence interval for σ is

$$\left(\sqrt{\frac{(n-1)s^2}{\chi^2_{1-\alpha/2,n-1}}}, \sqrt{\frac{(n-1)s^2}{\chi^2_{\alpha/2,n-1}}} \right)$$

- Tables for $\chi^2_{1-\alpha/2,n-1}$ are in Appendix A.3.
- χ^2 is not symmetric distribution, so $\chi^2_{1-\alpha,n} \neq -\chi^2_{\alpha,n}$

Example of interval estimation of σ^2

- Recall measurements of bat proximity to insect (in cm)

$$62, 52, 68, 23, 34, 45, 27, 42, 83, 56, 40$$

- Recall $n = 11$, $\bar{y} = 48.36$, and $s = 18.08$
- Take $\alpha = 0.05$, note that
  - $\chi^2_{0.05/2,11-1} = \chi^2_{0.025,10} = 3.247$
  - $\chi^2_{1-0.05/2,11-1} = \chi^2_{0.975,10} = 20.483$
- Then the 95% confidence interval for σ is

$$\left(\sqrt{\frac{10(18.08)^2}{20.483}}, \sqrt{\frac{10(18.08)^2}{3.247}}\right) = (12.63, 31.73)$$

- Note $s = 18.08$ *not* in center of this confidence interval.

Testing $H_0 : \sigma^2 = \sigma_0^2$

- Let s^2 denote the sample variance from n observations drawn from a normal distribution with unknown mean μ and unknown variance σ^2. Let $\chi^2 = \frac{(n-1)s^2}{\sigma_0^2}$.
  - To test $H_0 : \sigma^2 = \sigma_0^2$ versus $H_1 : \sigma^2 > \sigma_0^2$ at the α level of significance, reject H_0 if $\chi^2 \geq \chi^2_{1-\alpha,n-1}$.
  - To test $H_0 : \sigma^2 = \sigma_0^2$ versus $H_1 : \sigma^2 < \sigma_0^2$ at the α level of significance, reject H_0 if $\chi^2 \leq \chi^2_{\alpha,n-1}$.
  - To test $H_0 : \sigma^2 = \sigma_0^2$ versus $H_1 : \sigma^2 \neq \sigma_0^2$ at the α level of significance, reject H_0 if χ^2 is either (a) $\leq \chi^2_{\alpha/2,n-1}$ or (b) $\geq \chi^2_{1-\alpha/2,n-1}$.
- Note that we have limited our attention to t tests of Type I errors. It is possible to generalize tests for Type II errors, power curves, etc. for t tests.

Summary

Bruce M. Boghosian

Interval estimation
Hypothesis testing
Interval estimation of the variance
Hypothesis testing with the variance
Summary

- Normal random variables with unknown μ and σ.
- We can do interval estimation with such variables.
- We can do hypothesis testing with such variables.
- We compared t testing with more familiar z testing.
- We extended this methodology to confidence intervals.
- We can do hypothesis testing for the variance.