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

Bruce M. Boghosian

Department of Mathematics

Tufts University

Outline

Distribution of $\frac{(n-1)S^2}{\sigma^2}$

Bruce M. Boghosian

- Begin with n iid r.v.s in $N(\mu, \sigma)$, called $Y_1, \ldots, Y_n$.
- Standardize these variables to obtain $X_j = \frac{Y_j - \mu}{\sigma}$.
- Recall that we made an orthogonal transformation $\vec{Z} = A\vec{X}$
- The last row of the orthogonal matrix A was $\left(\frac{1}{\sqrt{n}}, \ldots, \frac{1}{\sqrt{n}}\right)$
- Hence $Z_n = \frac{1}{\sqrt{n}}(X_1 + \cdots + X_n) = \sqrt{n}\,\overline{X}$.
- And because the transformation is orthogonal, $\sum_j^n Z_j^2 = \sum_j^n X_j^2$

$$\sum_{j=1}^{n} Z_j^2 = \sum_{j=1}^{n-1} Z_j^2 + n\overline{X}^2 = \sum_{j}^{n} X_j^2 = \sum_{j=1}^{n}\left(X_j - \overline{X}\right)^2 + n\overline{X}^2$$

$$\therefore \quad \sum_{j=1}^{n-1} Z_j^2 = \sum_{j=1}^{n}\left(X_j - \overline{X}\right)^2 = \frac{1}{\sigma^2}\sum_{j=1}^{n}\left(Y_j - \overline{Y}\right)^2 = \frac{(n-1)S^2}{\sigma^2}$$

- Hence $\frac{(n-1)S^2}{\sigma^2}$ is distributed as a chi squared distribution with $n-1$ df.

Definition of sufficiency

- Recall the definition: Let $X_j = k_j$ for $j = 1, \ldots, n$ be a random sample of size n from $p_X(k;\theta)$. The statistic $\hat{\theta} = h(X_1, \ldots, X_n)$ is *sufficient* for θ if the likelihood function $L(\theta)$ factors into the product of the pdf for $\hat{\theta}$ and a constant that does not involve θ, i.e.,

$$L(\theta) = \prod_{j=1}^{n} p_X(k_j;\theta) = p_{\hat{\theta}}(\theta_e;\theta)\ b(k_1, \ldots, k_n).$$

- Similar statement for continuous random variables $Y_j = y_j$ for $j = 1, \ldots, n$,

$$L(\theta) = \prod_{j=1}^{n} f_Y(y_j;\theta) = f_{\hat{\theta}}(\theta_e;\theta)\ b(y_1, \ldots, y_n).$$

- We used this definition for Bernoulli and binomial random variables, e.g.

A more general factorization criterion

Bruce M. Boghosian

- **Thm:** Let $X_j = k_j$ for $j = 1, \ldots, n$ be a random sample of size n from the discrete pdf $p_X(k;\theta)$. The statistic $\hat{\theta}$ is *sufficient* for θ iff there are functions $g[h(k_1, \ldots, k_n);\theta]$ and $b(k_1, \ldots, k_n)$ such that

$$L(\theta) = \prod_{j=1}^{n} p_X(k_j;\theta) = g[h(k_1, \ldots, k_n);\theta]\ b(k_1, \ldots, k_n)$$

- Similar statement for continuous random variables $Y_j = y_j$ for $j = 1, \ldots, n$,

$$L(\theta) = \prod_{j=1}^{n} f_Y(y_j;\theta) = g[h(y_1, \ldots, y_n);\theta]\ b(y_1, \ldots, y_n)$$

A more general factorization criterion (continued)

- **Pf (for discrete case):** First suppose that $\hat{\theta}$ is sufficient. Then, by definition, we can write
$$L(\theta) = p_{\hat{\theta}}(\theta_e; \theta)\ b(k_1, \ldots, k_n),$$
which is of the desired form if we identify
$$g[h(k_1, \ldots, k_n); \theta] = p_{\hat{\theta}}(h(k_1, \ldots, k_n); \theta).$$
- Next suppose that
$$L(\theta) = \prod_{j=1}^{n} p_X(k_j; \theta) = g[h(k_1, \ldots, k_n); \theta]\ b(k_1, \ldots, k_n)$$
- We need to show that $g[h(k_1, \ldots, k_n); \theta]$ can always be rewritten in terms of the pdf of $\hat{\theta}$.

A more general factorization criterion (continued)

Bruce M. Boghosian

- Need to show $g[h(k_1,\ldots,k_n);\theta]$ can always be rewritten in terms of pdf of $\hat{\theta}$.
- Let c be some possible value of $h(k_1,\ldots,k_n)$, and let $A=\left\{\vec{k}\mid h(\vec{k})=c\right\}$ be the inverse image of c, so we write $A=h^{-1}(c)$. Then

$$\therefore\quad p_{\hat{\theta}}(c;\theta)=\sum_{\vec{k}\in A}p_{\vec{X}}(\vec{k})=\sum_{\vec{k}\in A}\prod_{j=1}^{n}p_{X_j}(k_j)=\sum_{\vec{k}\in A}g(c;\theta)b(\vec{k})=g(c;\theta)\left[\sum_{\vec{k}\in A}b(\vec{k})\right]$$

- It follows that

$$g(c;\theta)=\frac{p_{\hat{\theta}}(c;\theta)}{\sum_{\vec{k}\in A}b(\vec{k})}$$

$$\therefore\quad L(\theta)=p_{\hat{\theta}}(h(\vec{k});\theta)\frac{b(\vec{k})}{\sum_{\vec{k}\in A}b(\vec{k})}$$

and hence $\hat{\theta}$ is sufficient by definition, as was to be shown. $\square$

From Z tables to T tables

- The quantities z_α used in the tests described on the last slide are...
  - tabulated, e.g., in Appendix A.1 of the Larsen & Marx textbook.
  - given by intrinsic or library routines in many computer languages.
  - defined so area under the std. normal pdf to the right of $z = z_\alpha$ is equal to α.
- In like fashion, the quantities $t_{\alpha,n}$ are...
  - tabulated, e.g., in Appendix A.2 of the Larsen & Marx textbook.
  - given by intrinsic or library routines in many computer languages.
  - defined so area under the Student T_n pdf to the right of $t = t_{\alpha,n}$ is equal to α.
  - symmetric so that $P(T_n \leq -t_{\alpha,n}) = P(T_n \geq +t_{\alpha,n}) = \alpha$.

	α						
df	.20	.15	.10	.05	.025	.01	.005
1	1.376	1.963	3.078	6.3138	12.706	31.821	63.657
2	1.061	1.386	1.886	2.9200	4.3027	6.965	9.9248
3	0.978	1.250	1.638	2.3534	3.1825	4.541	5.8409
4	0.941	1.190	1.533	2.1318	2.7764	3.747	4.6041
5	0.920	1.156	1.476	2.0150	2.5706	3.365	4.0321
6	0.906	1.134	1.440	1.9432	2.4469	3.143	3.7074
⋮			⋮				
30	0.854	1.055	1.310	1.6973	2.0423	2.457	2.7500
∞	0.84	1.04	1.28	1.64	1.96	2.33	2.58

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

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

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

- Hence, if $y_1, \ldots, y_n$ is a random sample of size n from a normal distribution with known standard deviation σ and unknown mean 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

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

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 the T ratio: The one-sample t test

Bruce M. Boghosian

- 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

Bruce M. Boghosian

- Corrosion of metal pipe for underground cables in one year
- In the absence of plastic coating, average pit depth is 0.0042 inch
- $n = 10$ tests with the plastic coating yield the 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)

Bruce M. Boghosian

- H_0: The plastic coating has no effect, so $\mu = \mu_0 = 0.0042$ inch
- H_1: The plastic coating has a 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

Bruce M. Boghosian

- Note z tests can be used on data that is not normally distributed, as long as it has finite variance and n is sufficiently large that the 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 the quantity $\frac{\overline{Y}-\mu}{S/\sqrt{n}}$ 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 $f_{T_{n-1}}(t)$.
  - The t test is *robust with respect to departures from normality*, as is the z test.

Constructing confidence intervals for σ^2

Bruce M. Boghosian

- 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 chi squared distributed with $n-1$ df
- It follows that

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

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

Constructing confidence intervals for σ^2 (continued)

Bruce M. Boghosian

- It follows that 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, and follow the same conventions used for z and t tables; that is $P(\chi^2_n > \chi^2_{\alpha,n}) = \alpha$.
- NB: Chi squared is not a symmetric distribution, so $\chi^2_{1-\alpha,n} \neq -\chi^2_{\alpha,n}$

Example of interval estimation of σ^2

Bruce M. Boghosian

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

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

- Recall $n = 11$, $\overline{y} = 48.36$, 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 that $s = 18.08$ is *not* in the center of this confidence interval.

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

Bruce M. Boghosian

- 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

- We have learned to work with Gaussian random variables with unknown μ and σ.
- We have shown how to do interval estimation with such variables.
- We have shown how to do hypothesis testing with such variables.
- We have contrasted this t testing with z testing, with which we were already familiar.
- We have extended this to confidence intervals and hypothesis testing for the variance.