



Introduction to Statistical Inference

Guang Cheng, Evangelos Evangelou, Vered Madar and Jaeun Choi

May 20, 2008

The Undergraduate Workshop

Outline

I Theory

- * The Classical Problem of Statistical Inference
- * The Normal Model
 - Parameter Estimation (population mean and population scale)
 - Confidence Interval for the population mean
- * The Binomial Model

II MATLAB Examples

III Hypotheses Testing for the population mean - Normal Model

- * One-Sided and Two-Sided Tests and P-values

The Classical Problem

A finite set of *observations* (*a sample* if drawn randomly)

$$X_1, X_2, \dots, X_n$$

reflecting some general population.

The Classical Problem

A finite set of *observations* (*a sample* if drawn randomly)

$$X_1, X_2, \dots, X_n$$

reflecting some general population.

Sample: Heights(ft) of 17 5th grade kids in Miss Kenny class at Durham School:

4.5, 4.6, 4.9, 5.1, 4.2, 5.1, 5.5, 5.0, 4.8, 5.2, 4.5, 4.8, 4.7, 5.2, 6.1, 5.4, 4.8.

Population: Heights of all 5th grade boys in USA.

The Classical Problem

A finite set of *observations* (*a sample* if drawn randomly)

$$X_1, X_2, \dots, X_n$$

reflecting some general population.

Sample: Values of 0s and 1s, for each student that fails(0) or passes(1) the EOG math:

0, 1, 0, 0, 1, 1, 0, 1, 1, 1, 1, 0, 1, 1, 0, 1, 1.

Population: Success of all 5th grade students in USA that passes math EOG.

The Classical Problem - Assumptions

A finite set of *observations*

$$X_1, X_2, \dots, X_n$$

The Classical Problem - Assumptions

A finite set of *observations*

$$X_1, X_2, \dots, X_n$$

- * **Independence:** Each observation (student,height), X_i , is independent of all others;

The Classical Problem - Assumptions

A finite set of *observations*

$$X_1, X_2, \dots, X_n$$

- * **Independence:** Each observation (student,height), X_i , is independent of all others;
- * **Model:** Assume a model with parameters and an error term ϵ_i for the population

The Classical Problem - Assumptions

A finite set of *observations*

$$X_1, X_2, \dots, X_n$$

- * **Independence:** Each observation (student,height), X_i , is independent of all others;
- * **Model:** Assume a model with parameters and an error term ϵ_i for the population
- * **Distribution:** We know the asymptotic or the exact distribution or behavior of the error term. (e.g., normal,binomial,...)

The Classical Problem - Assumptions

A finite set of *observations*

$$X_1, X_2, \dots, X_n$$

- * **Independence:** Each observation (student,height), X_i , is independent of all others;
- * **Model:** Assume a model with parameters and an error term ϵ_i for the population
- * **Distribution:** We know the asymptotic or the exact distribution or behavior of the error term. (e.g., normal,binomial,...)

Last two assumptions - needed to be checked!!!

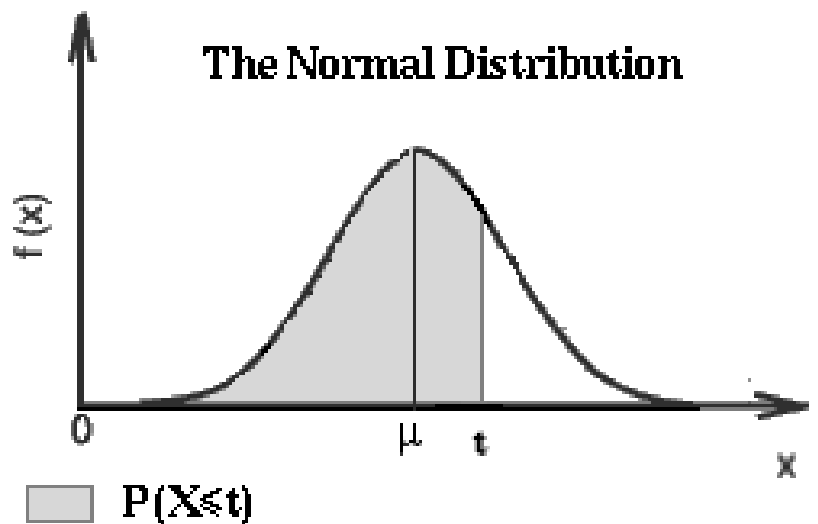
The Normal Distribution

The random variable X is normally distributed, $X \sim \mathcal{N}(\mu, \sigma^2)$, if

$$\Phi(x) \equiv \Pr(X \leq x) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^x \exp^{-\frac{(t-\mu)^2}{2\sigma^2}} dt,$$

where

- * μ - the population mean (any real value);
- * σ^2 - the population standard deviation ($\sigma^2 > 0$).



Estimation of μ

The Sample Mean

$$\bar{X} = \frac{1}{n} (X_1 + X_2 + \cdots + X_n).$$

- * For n large enough, the distribution of \bar{X} is approximately normal $\mathcal{N}(\mu, \sigma^2/n)$.
- * “On average” the arithmetic mean is the population mean $E(\bar{X}) = E(X_1) = \mu$;
- * In the Heights example: $\bar{X} = 4.96$, but true population mean is 4.8 ft.

Estimation of σ^2

The Sample Variance:

$$S^2 = \frac{1}{n-1} ((X_1 - \bar{X})^2 + (X_2 - \bar{X})^2 + \cdots + (X_n - \bar{X})^2)$$

Why S^2 ?

- * “On average” S^2 is σ^2 , that is $E(S^2) = \sigma^2$;
- * In the Heights example: $S^2 = 0.20 = (0.45)^2$.

A $1 - \alpha$ Confidence Intervals for μ

For a predetermined probability $\alpha \in (0, 1)$ (say, $\alpha = 0.05 = 5\%$) we would like to form an interval

$$[\bar{X} - d, \bar{X} + d]$$

in a way that the true population mean (μ) is inside with probability:

$$\Pr(\bar{X} - d \leq \mu \leq \bar{X} + d) \geq 1 - \alpha.$$

A $1 - \alpha$ Confidence Intervals for μ

For a predetermined probability $\alpha \in (0, 1)$ (say, $\alpha = 0.05 = 5\%$) we would like to form an interval

$$[\bar{X} - d, \bar{X} + d]$$

in a way that the true population mean (μ) is inside with probability:

$$\Pr(\bar{X} - d \leq \mu \leq \bar{X} + d) \geq 1 - \alpha.$$

This is a $1 - \alpha$ *confidence interval* and $1 - \alpha$ is the *confidence level*.

A $1 - \alpha$ Confidence Intervals for μ — cont.

- * We don't know μ or σ^2 ;
- * So we estimate μ by $\bar{X} = 4.96$ and σ by $S = 0.45$.

A $1 - \alpha$ Confidence Intervals for μ — cont.

- * We don't know μ or σ^2 ;
- * So we estimate μ by $\bar{X} = 4.96$ and σ by $S = 0.45$.
- * We use the fact that

$$T = \frac{\bar{X}}{\frac{S}{\sqrt{n}}}$$

is t distributed with mean μ and standard deviation $\sigma^2 = 1$ and $n - 1$ degrees of freedom.

- * The t distribution behaves like the Normal distribution for large sample size n .

A $1 - \alpha$ Confidence Intervals for μ — cont.

- * We don't know μ or σ^2 ;
- * So we estimate μ by $\bar{X} = 4.96$ and σ by $S = 0.45$.
- * We use the fact that

$$T = \frac{\bar{X}}{\frac{S}{\sqrt{n}}}$$

is t distributed with mean μ and standard deviation $\sigma^2 = 1$ and $n - 1$ degrees of freedom.

- * The t distribution behaves like the Normal distribution for large sample size n .
- * A $1 - \alpha$ confidence interval for the mean will be

$$\left[\bar{X} - \frac{S}{\sqrt{n}} \cdot q, \bar{X} + \frac{S}{\sqrt{n}} \cdot q \right],$$

q is the min value that satisfies $\Pr(T \geq q) = \alpha/2$ and $\Pr(T \leq -q) = \alpha/2$.

A $1 - \alpha$ Confidence Intervals for μ ---cont.

In the heights example: A 95% confidence interval is

$$\left[\bar{X} - q \cdot \frac{S}{\sqrt{n}}, \bar{X} + q \cdot \frac{S}{\sqrt{n}} \right]$$

we take $q = 2.12$ (MATLAB: `tinv(0.025,16)`) and $\frac{S}{\sqrt{n}} = 0.11$, so we get

$$[4.96 \pm 2.12 \cdot 0.11] = [4.73, 5.20] \text{ ft.}$$

With confidence level of 95% we say that the population mean height is between 4.73 and 5.2 ft.

The Binomial (Bernoulli) Model

There are n independent observations X_1, X_2, \dots, X_n , each can be either 0 or 1.

One parameter p - the probability of success

$$\Pr(X_i = 0) = 1 - p \quad \text{and} \quad \Pr(X_i = 1) = p.$$

The Binomial (Bernoulli) Model

There are n independent observations X_1, X_2, \dots, X_n , each can be either 0 or 1.

One parameter p - the probability of success

$$\Pr(X_i = 0) = 1 - p \quad \text{and} \quad \Pr(X_i = 1) = p.$$

* The Sample Mean:

$$\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i \quad (\text{In the EOG example: } = \frac{11}{17} = 0.647);$$

* “On average” $E(\bar{X}) = p$;

* The Sample Variance: $S^2 = \bar{X}(1 - \bar{X}) (= 0.647 \cdot (1 - 0.647));$

Confidence Interval for p

* A $1 - \alpha$ confidence interval for p :

$$\left[\bar{X} - q \cdot \frac{S}{\sqrt{n}}, \bar{X} + q \cdot \frac{S}{\sqrt{n}} \right].$$

Confidence Interval for p

- * A $1 - \alpha$ confidence interval for p :

$$\left[\bar{X} - q \cdot \frac{S}{\sqrt{n}}, \bar{X} + q \cdot \frac{S}{\sqrt{n}} \right].$$

- * $\bar{X} = 0.647$, and $\frac{S}{\sqrt{n}} = 0.477$;
- * We compute q as for a Normal distribution (this is a common approximation).
- * For a 95% confidence interval, $q = 1.96$ (MATLAB: `tinu(0.025,100)`)

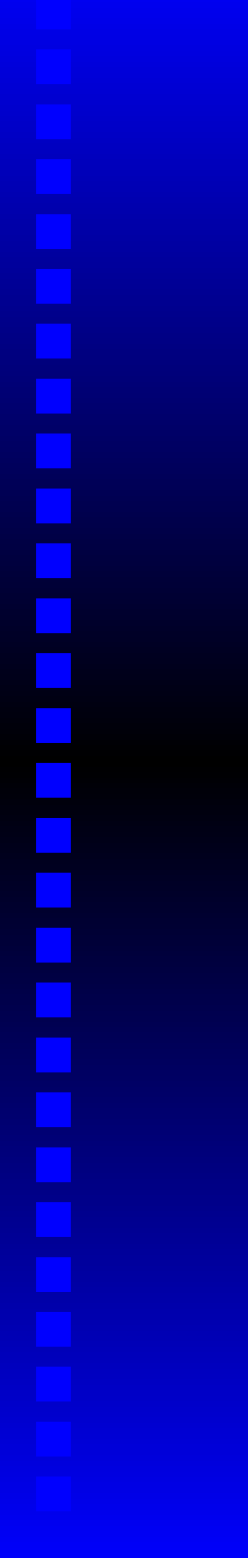
$$[0.647 \pm 1.96 \cdot 0.477] = [0.427, 0.874].$$

Part II - Data Sets for MATLAB examples

- * **pH data** The data consist of acidity levels (pH) in 105 samples of rainwater. pH ranges from 0 to 14 and distilled water has pH 7. As the water becomes more acid, the pH goes down. The pH of water is important to environmentalists because of the problem of acid rain. (From Moore & McCabe, Introduction to the practice of statistics)
- * **Salmon data** Salmon is born in freshwater rivers and streams and then swim out to the ocean. Researches have studied the growth of freshwater salmon and first year salmon caught in the ocean by measuring the radius of their growth rings. (From Johnson & Tsui, Statistical reasoning and methods)
- * **Light data** This dataset is the result of the classic study conducted by Michelson on the speed of light in air in 1879. The response variable is speed of light (in millions of meters per second). The data was included as part of a larger study by Dorsey, Ernest N. (1944) on the velocity of light as reported in the Transactions of the American Philosophical Society. (From statistical reference datasets)
- * **IQ data:** Displays the IQ scores of 60 fifth grade students. (From Moore & McCabe, Introduction to the practice of statistics)
- * **Oil data** How much oil in the wells in a given filed will produce is key information in deciding whether to drill more wells. This dataset contains the amount of oil recovered from 64 wells in the Devonian Richmond Dolomite area of the Michigan basin, in thousands of barrels. (From Moore & McCabe, Introduction to the practice of statistics)

Data Sets for MATLAB examples - cont.

- * **Emissions data** This dataset contains the amounts, in grams per mile, of three pollutants (HydroCarbon, Carbon Monoxide and Nitrogen Oxides) in the exhaust of 46 vehicles of the same type measured under standard conditions prescribed by the EPA. (From Moore & McCabe, Introduction to the practice of statistics)
- * **Test data** The dataset gives the pretest and posttest scores in Spanish for 20 high school Spanish teachers who attended an intensive summer course in Spanish. (From Moore & McCabe, Introduction to the practice of statistics)
- * **Fitness data** Data on students taking a course designed to introduce them to a variety of training techniques. The variables are: gender (1=male, 2=female), pretest body fat, posttest body fat, pretest time to run 1.5 miles (seconds), posttest to run 1.5 miles (seconds), pretest to row 2.5Km (seconds), posttest to row 2.5Km (seconds), pretest number of sit-ups completed in 1 minute, posttest number of sit-ups completed in 1 minute. (From Johnson & Tsui, Statistical reasoning and methods)



Part III : Test of Hypotheses

Test of Hypotheses

- * A hypothesis test is a way to decide whether the data strongly support point of view or another.
- * All involve:
 - * A null hypothesis H_0 (we intend to reject) and an alternative hypothesis H_1
 - * A level of test (α) - [small enough: 1% or 5%]
 - * A test statistic - to reflect our data

Step 1: Pick the null and alternative hypotheses

- cont'

Possible Null and Alternative Hypotheses

- * **One Sided Test:** $H_0 : \theta \geq \theta_0$ vs $H_1 : \theta < \theta_0$ or ($H_0 : \theta \leq \theta_0$ vs $H_1 : \theta > \theta_0$)

Step 1: Pick the null and alternative hypotheses

- cont'

Possible Null and Alternative Hypotheses

* **One Sided Test:** $H_0 : \theta \geq \theta_0$ vs $H_1 : \theta < \theta_0$ or ($H_0 : \theta \leq \theta_0$ vs $H_1 : \theta > \theta_0$)

Suppose we wish to claim that the mean height > 4.8

H_0 : Population mean height of 5th grade boy is ≤ 4.8 ft.

against

H_1 : Population mean height of 5th grade boy is > 4.8 ft.

Step 1: Pick the null and alternative hypotheses

- cont'

Possible Null and Alternative Hypotheses

- * **One Sided Test:** $H_0 : \theta \geq \theta_0$ vs $H_1 : \theta < \theta_0$ or ($H_0 : \theta \leq \theta_0$ vs $H_1 : \theta > \theta_0$)

Suppose we wish to claim that the mean height > 4.8

H_0 : Population mean height of 5th grade boy is ≤ 4.8 ft.

against

H_1 : Population mean height of 5th grade boy is > 4.8 ft.

- * **Two Sided Test:** $H_0 : \theta = \theta_0$ vs $H_1 : \theta \neq \theta_0$

H_0 : Mean of annual global temperature = $56 F^\circ$

against

H_1 : Mean of annual global temperature $\neq 56 F^\circ$.

Step 2 and Step 3

Step 2. Use the test statistic:

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

Under H_0 the test statistic is

$$T \sim \text{approx. Normal } \mathcal{N}(\mu = 0, \sigma^2 = 1)$$

Compute t - the value of T obtained for your sample.

Step 2 and Step 3

Step 2. Use the test statistic:

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

Under H_0 the test statistic is

$$T \sim \text{approx. Normal } \mathcal{N}(\mu = 0, \sigma^2 = 1)$$

Compute t - the value of T obtained for your sample.

Step 3. Apply P-value:

P-value is the chance that our observed data (reflected by t) supports the alternative when H_0 is actually true.

- * **One-sided P-value** $\Pr(T \geq t)$ (or $\Pr(T \leq t)$)
- * **Two-sided P-value** $\Pr(|T| \geq |t|)$.

Step 2 and Step 3

Step 2. Use the test statistic:

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

Under H_0 the test statistic is

$$T \sim \text{approx. Normal } \mathcal{N}(\mu = 0, \sigma^2 = 1)$$

Compute t - the value of T obtained for your sample.

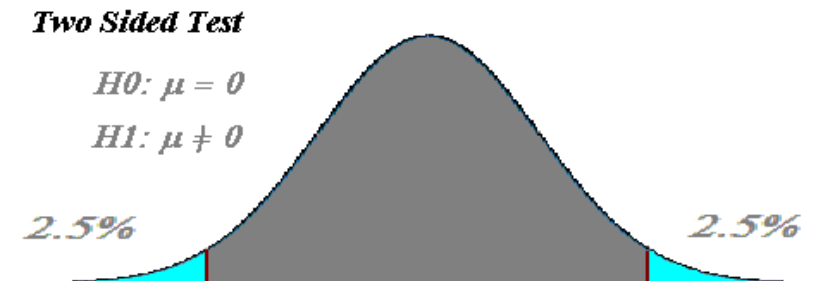
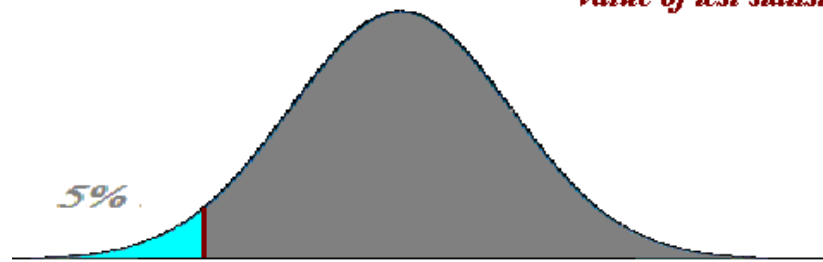
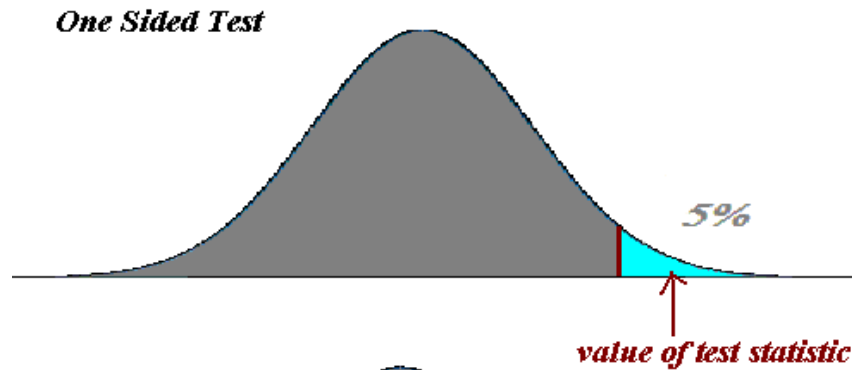
Step 3. Apply P-value:

P-value is the chance that our observed data (reflected by t) supports the alternative when H_0 is actually true.

- * One-sided P-value $\Pr(T \geq t)$ (or $\Pr(T \leq t)$)
- * Two-sided P-value $\Pr(|T| \geq |t|)$.

P-value $< \alpha$ (such as $< 5\%$ or $< 1\%$) \implies Reject H_0 , otherwise **do not reject H_0 .**

The Usage of P-values in Hypotheses Testing



Example of P-value

New drug may extend life expectancy. Data of 36 random people taking this drug yield a mean lifespan of 76 years and a standard deviation 12 years. If we know that the average US lifespan is 72, can we show that the new drug extends life expectancy significantly?

Example of P-value

New drug may extend life expectancy. Data of 36 random people taking this drug yield a mean lifespan of 76 years and a standard deviation 12 years. If we know that the average US lifespan is 72, can we show that the new drug extends life expectancy significantly?

Step 1:

H_0 : The mean of lifetime with the new drug \leq 72 years.

H_1 : The mean of lifetime with the new drug $>$ 72 years.

Example of P-value

New drug may extend life expectancy. Data of 36 random people taking this drug yield a mean lifespan of 76 years and a standard deviation 12 years. If we know that the average US lifespan is 72, can we show that the new drug extends life expectancy significantly?

Step 1:

H_0 : The mean of lifetime with the new drug ≤ 72 years.

H_1 : The mean of lifetime with the new drug > 72 years.

Step 2: Compute the test statistic

$$t = \frac{\bar{X} - \mu_0}{S/\sqrt{n}} = \frac{76 - 72}{12/\sqrt{36}} = 2$$

Example of P-value

New drug may extend life expectancy. Data of 36 random people taking this drug yield a mean lifespan of 76 years and a standard deviation 12 years. If we know that the average US lifespan is 72, can we show that the new drug extends life expectancy significantly?

Step 1:

H_0 : The mean of lifetime with the new drug ≤ 72 years.

H_1 : The mean of lifetime with the new drug > 72 years.

Step 2: Compute the test statistic

$$t = \frac{\bar{X} - \mu_0}{S/\sqrt{n}} = \frac{76 - 72}{12/\sqrt{36}} = 2$$

Step 3: find the p-value

$$P\text{value} = \Pr(T > 2) = 0.0267 \quad \text{MATLAB: } 1 - \text{tcdf}(2, 35)$$

Example of P-value

New drug may extend life expectancy. Data of 36 random people taking this drug yield a mean lifespan of 76 years and a standard deviation 12 years. If we know that the average US lifespan is 72, can we show that the new drug extends life expectancy significantly?

Step 1:

H_0 : The mean of lifetime with the new drug ≤ 72 years.

H_1 : The mean of lifetime with the new drug > 72 years.

Step 2: Compute the test statistic

$$t = \frac{\bar{X} - \mu_0}{S/\sqrt{n}} = \frac{76 - 72}{12/\sqrt{36}} = 2$$

Step 3: find the p-value

$$P\text{value} = \Pr(T > 2) = 0.0267 \quad \text{MATLAB: } 1 - \text{tcdf}(2, 35)$$

Interpretation: P-value close to zero — reject H_0 - the new drug extends life significantly in a 5% test.



Thanks

That's All Folks!