Statistics for Engineers Lecture 5
Statistical Inference

Chong Ma

Department of Statistics
University of South Carolina
chongm@email.sc.edu

March 13, 2017

Outline

1. Populations and Samples
2. Parameters and Statistics
3. Point Estimator and Sampling Distribution
4. Central Limit Theorem
5. t-distribution

Populations and Samples

- **Statistical Inference** deals with making (probabilistic) statements about a population of individuals based on information that is contained in a sample taken from the population. We do this by
  - (a) estimating unknown population parameters with sample statistics.
  - (b) quantifying the uncertainty(variability) that arises in the estimation process.
- **Population**: The total set of subjects in which we are interested such as "All undergraduates at USC", "every atom composing a crystal", batteries, about which we would like to make a statement(e.g., median IQ score, mean size, mean lifetime).
- **Sample**: The subset of the population for whom we have data, often random selected. Mathematically, it means that all observations are independent and follow the same probability distribution. Informally, this means that each sample(of the same size) has the same chance of being selected.

Populations and Samples

Remarks: It is generally accepted that the entire population can not be measured(b/c too large or too time-consuming to do so). A random sample is usually the best way to obtain individuals that are "representative" of the entire population. Denote a random sample of observations by

$$Y_1, Y_2, \ldots, Y_n$$

Where

- Y_i is the value of Y for for the ith individual in the sample.
- n is the sample size that indicates how many individuals are in the sample.
- Lower case notation $y_1, y_2, \ldots, y_n$ is used when citing numerical values.

Population and Samples

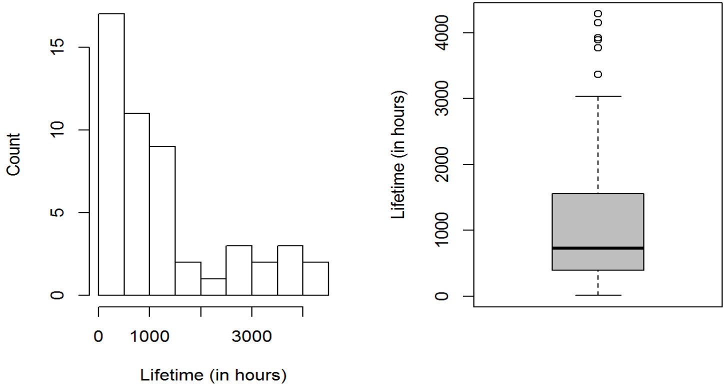

Figure 1: Histogram(left) and boxplot(right) of the battery lifetime data (measured in hours).

Population and Samples

Consider the following random sample of $n = 50$ battery lifetimes $y_1, y_2, \ldots, y_{50}$ measured in hours.

4258	2066	2584	1009	318	1429	981	1402
...	...	...	...	...	...	...	...
99	510	582	308	3367	99	373	454

- Which continuous probability distribution seems to display the same type of pattern in the histogram?
- An exponential(λ) model seems reasonable here. What is λ?
- λ is called a (population) **parameter**. It describes the distribution which is used to model the entire population of batteries.
- In general, (population) **parameters** which characterize probability distributions are unknown.

Outline

1. Populations and Samples
2. Parameters and Statistics
3. Point Estimator and Sampling Distribution
4. Central Limit Theorem
5. t-distribution

Parameters and Statistics

- **Parameter**: A numerical summary of the population, such as a population proportion *p* for a categorical variable fixed but usually unknown.
- **Statistic**: A numerical summary of a sample taken from the population, such as the sample mean, sample proportion, sample median and so on.

All of the probability distributions that we talked about in chapters 3-5 were characterized by population(model) parameters. For example,

- $\mathcal{N}(\mu, \sigma^2)$ characterized by the population mean μ and population variance σ^2.
- $\text{Poisson}(\lambda)$ characterized by one parameter, the population mean λ.
- $\text{Weibull}(\beta, \eta)$ characterized by the shape parameter β and the scale parameter η.

Parameters and Statistics

Suppose that $Y_1, Y_2, \ldots, Y_n$ is a random sample from a population. The **sample mean** is

$$\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$$

The **sample variance** is

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \bar{Y})^2$$

The **sample standard deviation** is the positive square root of the sample variance,

$$S = \sqrt{S^2} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \bar{Y})^2}$$

Paramters and Statistics

The table below succinctly summarizes the differences between a population and a sample (a parameter and a statistic):

Group of individuals	Numerical quantity	Status
Population(not observed)	Parameter	Unknown
Sample(observed)	Statistic	Calculated from sample data

In the battery lifetime example,

- $\bar{y} = 1274.14$ is an **estimate** of the population mean μ.
- $s^2 = 1505156$ is an **estimate** of the population variance σ^2.
- $s = 1226.848$ is an **estimate** of the population standard deviation σ.

Outline

1. Populations and Samples
2. Parameters and Statistics
3. Point Estimator and Sampling Distribution
4. Central Limit Theorem
5. t-distribution

Point Estimators

For notational simplicity, we denote population parameters by θ(a "wild card"). θ could denote a population mean, population variance or a Weibull or Gamma model parameters. A **point estimator** $\hat{\theta}$ is a statistic that is used to estimate a population parameter θ. Common examples of point estimators are

$$\begin{aligned} \bar{Y} &\rightarrow \text{ a point estimator for } \mu\text{(popultaion mean)} \\ S^2 &\rightarrow \text{ a point estimator for } \sigma^2\text{(popultaion mean)} \\ S &\rightarrow \text{ a point estimator for } \sigma\text{(popultaion mean)} \end{aligned}$$

Remarks: A point estimator $\hat{\theta}$ is a statistic, so it depends on the sample of data $Y_1, Y_2, \ldots, Y_n$.

- The data $Y_1, Y_2, \ldots, Y_n$ comes from the sampling process, that is, different random samples yield different data sets $Y_1, Y_2, \ldots, Y_n$.
- In this light, because the sample values $Y_1, Y_2, \ldots, Y_n$ will vary from sample to sample, so will value of $\hat{\theta}$. It therefore makes perfect sense to think about the **distribution** of $\hat{\theta}$ itself.

Point Estimator

The distribution of $\hat{\theta}$ is called its **sampling distribution**. A sampling distribution describes how the estimator $\hat{\theta}$ varies in repeated sampling.

Accuracy $\hat{\theta}$ is an **unbiased estimator** of θ if and only if $E(\hat{\theta}) = \theta$. Unbiasedness is a characteristic describing the center of a sampling distribution, which deals with **accuracy**.

Precision The **standard error** of a point estimator $\hat{\theta}$ is equal to

$$\mathrm{se}(\hat{\theta}) = \sqrt{\mathrm{var}(\hat{\theta})}$$

An estimator's standard error measures the amount of variability in the point estimator $\hat{\theta}$. Therefore,

$$\text{smaller } \mathrm{se}(\hat{\theta}) \Leftrightarrow \hat{\theta} \text{ more precise}$$

Sampling Distribution

Suppose that $Y_1, Y_2, \ldots, Y_n$ is a random sample from a $N(\mu, \sigma^2)$ distribution. The sample mean $\bar{Y}$ follows the **sampling distribution**:

$$\bar{Y} \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right)$$

where it indicates that

- $E(\bar{Y}) = \mu$, the sample mean $\bar{Y}$ is an **unbiased estimator** of the population mean μ.
- $\mathrm{var}(\bar{Y}) = \frac{\sigma^2}{n}$ is the variance of $\bar{Y}$.

Example Assume the distribution of

$$Y = \text{time(in seconds) to react to brake lights during in-traffic driving}$$

follows the distribution $Y \sim \mathcal{N}(\mu = 1.5, \sigma^2 = 0.16)$. We call this the **population distribution**, because it describes the distribution of values of Y for all individuals in the population (in-traffic drivers).

Sampling Distribution

(a) Suppose that we take a random sample of $n = 5$ drivers from the population with times $Y_1, Y_2, \ldots, Y_5$. What is the distribution of the sample mean $\bar{Y}$?

Sampling Distribution

(a) Suppose that we take a random sample of $n = 5$ drivers from the population with times $Y_1, Y_2, \ldots, Y_5$. What is the distribution of the sample mean $\bar{Y}$?
Note that the sample size $n = 5$, $\mu = 1.5$ and $\sigma^2 = 0.16$, we have

$$\bar{Y} \sim \mathcal{N}(\mu, \frac{\sigma^2}{n}) \Rightarrow \bar{Y} \sim \mathcal{N}(1.5, 0.032)$$

Sampling Distribution

(a) Suppose that we take a random sample of $n = 5$ drivers from the population with times $Y_1, Y_2, \ldots, Y_5$. What is the distribution of the sample mean $\bar{Y}$?
Note that the sample size $n = 5$, $\mu = 1.5$ and $\sigma^2 = 0.16$, we have

$$\bar{Y} \sim \mathcal{N}(\mu, \frac{\sigma^2}{n}) \Rightarrow \bar{Y} \sim \mathcal{N}(1.5, 0.032)$$

(b) Suppose that we take a random sample of $n = 25$ drivers from the population with times $Y_1, Y_2, \ldots, Y_{25}$. What is the distribution of the sample mean $\bar{Y}$?

Sampling Distribution

(a) Suppose that we take a random sample of $n = 5$ drivers from the population with times $Y_1, Y_2, \ldots, Y_5$. What is the distribution of the sample mean $\bar{Y}$?
Note that the sample size $n = 5$, $\mu = 1.5$ and $\sigma^2 = 0.16$, we have

$$\bar{Y} \sim \mathcal{N}(\mu, \frac{\sigma^2}{n}) \Rightarrow \bar{Y} \sim \mathcal{N}(1.5, 0.032)$$

(b) Suppose that we take a random sample of $n = 25$ drivers from the population with times $Y_1, Y_2, \ldots, Y_{25}$. What is the distribution of the sample mean $\bar{Y}$?
Note that the sample size $n = 25$, $\mu = 1.5$ and $\sigma^2 = 0.16$, we have

$$\bar{Y} \sim \mathcal{N}(\mu, \frac{\sigma^2}{n}) \Rightarrow \bar{Y} \sim \mathcal{N}(1.5, 0.0064)$$

Sampling Distribution

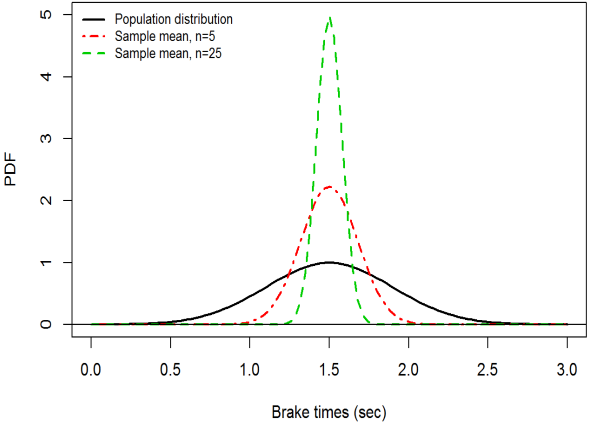

Figure 2: Brake time. Population distribution $Y \sim \mathcal{N}(\mu = 1.5, \sigma^2 = 0.16)$. Also depicted are the sampling distribution of $\bar{Y}$ when $n = 5$ and $n = 15$.

Outline

1 Populations and Samples

2 Parameters and Statistics

3 Point Estimator and Sampling Distribution

4 Central Limit Theorem

5 t-distribution

Central Limit Theorem(CLT)

Given certain conditions, the arithmetic mean of a sufficiently large number of independent random variables, each with a well-defined(finite) expected value(μ) and finite variance(σ^2), will be approximately normally distributed, regardless of the underlying distribution. Mathematically, it can be rewritten as follows.

CLT

Suppose $\{Y_1, Y_2, \ldots, Y_n\}$ is a sequence of i.i.d random variables with $E[Y_i] = \mu$ and $Var(Y_i) = \sigma^2 < \infty$. Then as n approaches infinity, the random variable $\frac{\sqrt{n}(\bar{Y}_n - \mu)}{\sigma}$ converge in distribution to the standard normal distribution $N(0, 1)$.

In other words,

$$\bar{Y}_n \sim \mathcal{AN}(\mu, \frac{\sigma^2}{n})$$

Central Limit Theorem

Example The time to death for rats injected with a toxic substance, denoted by Y (measured in days), follows an exponential distribution with $\lambda = 1/5$. That is, the **population distribution** is

$$Y \sim \text{exponential}(\lambda = 1/5)$$

Suppose that $n = 25$ rats are injected with the toxic substance. What is the probability the sample mean survival time $\bar{Y}$ will greater than 7 days?

Central Limit Theorem

Example The time to death for rats injected with a toxic substance, denoted by Y (measured in days), follows an exponential distribution with $\lambda = 1/5$. That is, the **population distribution** is

$$Y \sim \text{exponential}(\lambda = 1/5)$$

Suppose that $n = 25$ rats are injected with the toxic substance. What is the probability the sample mean survival time $\bar{Y}$ will greater than 7 days? Note that $n = 25$, $\mu = \frac{1}{\lambda} = 5$ and $\sigma^2 = \frac{1}{\lambda^2} = 25$, the CLT says that

$$\bar{Y} \sim \mathcal{AN}(\mu, \frac{\sigma^2}{n}) \Rightarrow \bar{Y} \sim \mathcal{AN}(5, 1)$$

Therefore,

$$\begin{aligned} P(\bar{Y} \geq 7) &= 1 - P(\bar{Y} < 7) \\ &= 1 - pnorm(7, 5, 1) \\ &= 0.023 \end{aligned}$$

Central Limit Theorem

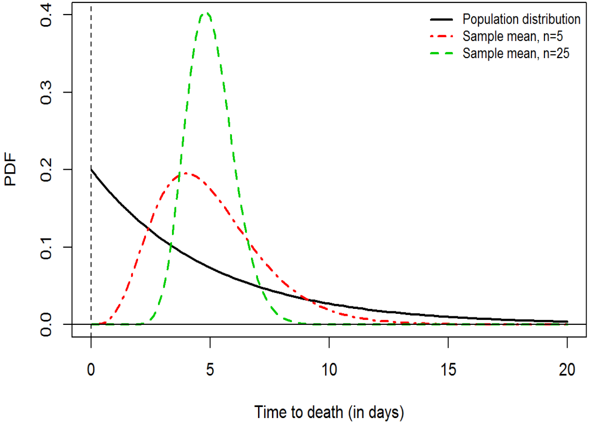

Figure 3: Rat death time. Population distribution $Y \sim \text{exponential}(\lambda = 1/5)$. Also depicted are the sampling distributions of $\bar{Y}$ when $n = 5$ and $n = 25$.

Outline

1. Populations and Samples
2. Parameters and Statistics
3. Point Estimator and Sampling Distribution
4. Central Limit Theorem
5. t-distribution

t distribution

Suppose that $Y_1, Y_2, \dots, Y_n$ is a random sample from a $\mathcal{N}(\mu, \sigma^2)$ distribution. Recall previous slides, we know that the sample mean $\bar{Y}$ has the following distribution:

$$\bar{Y} \sim \mathcal{N}(\mu, \frac{\sigma^2}{n})$$

If we standardize $\bar{Y}$, we obtain

$$Z = \frac{\bar{Y} - \mu}{\sigma/\sqrt{n}} \sim \mathcal{N}(0, 1)$$

If we replace the population standard deviation σ with the sample standard deviation S, we get a new sampling distribution

$$t = \frac{\bar{Y} - \mu}{S/\sqrt{n}} \sim t_{n-1}$$

a **t distribution** with degrees of freedom $\nu = n - 1$.

t distribution

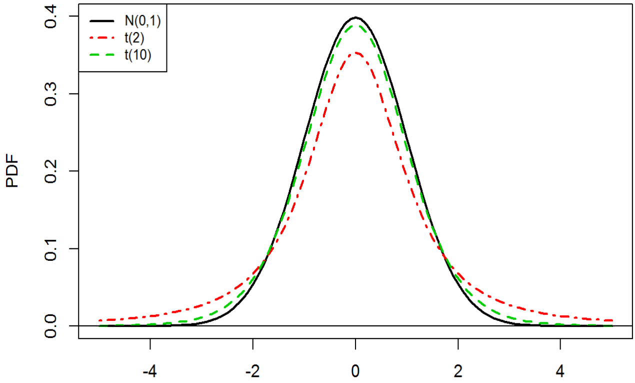

Figure 4: Probability distribution of $\mathcal{N}(0, 1)$, t_2 and t_{10}.

t distribution

The t distribution has following characteristics:

- It is continuous and symmetric about 0(just like the standard normal pdf).
- It is indexed by a value ν callsed the **degrees of freedom**. In practice, ν is often an integer(related to sample size).
- As $\nu \to \infty$, $t_\nu \to \mathcal{N}(0, 1)$.
- When compared to the standard normal pdf, the t pdf generally is less peaked and has more probability (area) in the tails.

The pdf for t_ν is

$$f(t) = \frac{\Gamma(\frac{\nu+1}{2})}{\sqrt{\nu\pi}\Gamma(\frac{\nu}{2})} \left(1 + \frac{t^2}{\nu}\right)^{-\frac{\nu+1}{2}}$$

In R, we use the code **pt(t,ν)** for computing the cdf $F_T(t)$ and **qt(t,ν)** for computing the quantile ϕ_p.

t distribution

Example Hollow pipes are to be used in an electrical wiring project. In testing "1-inch" pipes, the data collected by a design engineer. The data are measurements of Y, the **outside diameter** of this type of pipe (measured in inches). These $n = 25$ pipes were randomly selected and measured-all in the same location. The manufactures of this pipe claim that the population distribution is normal(Gaussian) and that the mean outside diameter is $\mu = 1.29$ inches. **Under this assumption** (which may or may not be true), calculate the value of

$$t = \frac{\bar{y} - \mu}{s/\sqrt{n}}$$

You can find the data in R tutorial. The sample mean and sample standard deviation are $\bar{Y} = 1.299$ and $s = 0.011$, therefore

$$t = \frac{\bar{y} - \mu}{s/\sqrt{n}} = \frac{1.299 - 1.29}{0.011/\sqrt{25}} \approx 4.096$$

t distribution

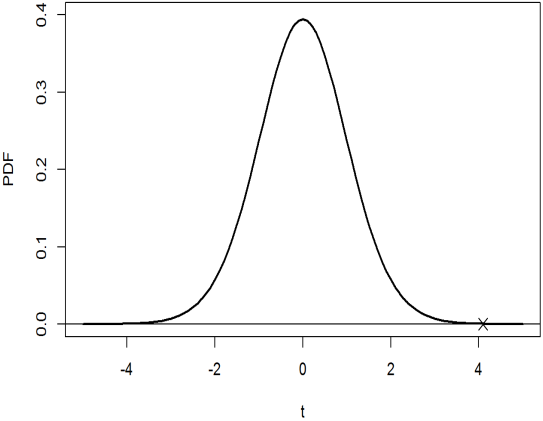

Figure 5: t_{24} probability density function. An "×" indicates that $t = 4.096$.

Normal QQ plot

Remarks: The t distribution result still approximately holds, even if the underlying population distribution is not perfectly normal. We also say the sampling distribution of t is robust to the normality assumption. The approximation is best when

- the sample size is large($n \geq 30$).
- the population distribution is more symmetric(not highly skewed).

Recall that we used Weibull qq plot to assess the Weibull model assumption in the last lecture, we can use a normal **quantile-quantile(qq) plot** to assess the normality assumption. The plot is constructed as follows:

- On the vertical axis, we plot the ascending-ordered observed data.
- On the horizontal axis, we plot the ordered theoretical quantiles from the distribution(model) assumed for the observed data(here, normal).

Normal QQ plot

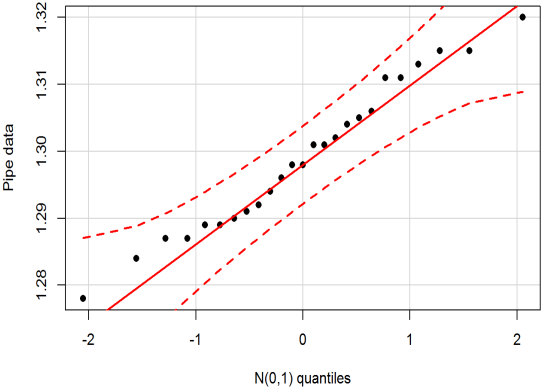

Figure 6: Pipe diameter data. Normal QQ plot. The observed data are plotted versus the theoretical quantiles from a standard normal distribution. The line added passes through the first and third quantiles.