Chapter 7, 8-Estimation and confidence intervals

After we have selected a sample, we know the responses of the individuals in the sample. The usual reason for taking a sample is to to infer from the sample data some conclusion about the wider population that the sample represents.

1 Chapter 7: Point estimation

In statistics, point estimation involves the use of sample data to calculate a single value (known as a point estimate or statistic) which is to serve as a "best guess" or "best estimate" of an unknown population distribution parameter.

- Based on the data of flipping a coin 10 times, we calculate the proportion of "heads" and use it to estimate the probability of "head" for the coin.
- Based on the data of 50 randomly selected UNM first-year college students, we calculate the average age and use it to estimate the average age of all UNM first-year students.
- Based on the data of 10 randomly selected days of traffic data, we calculate the average number of vehicles passing Lomas and University intersection from 6-8am. Then we use the average divided by 2 to estimate the rate parameter of a Poisson distribution that may describe the traffic.
- Based on the data of 20 experiment data on treatment and placebo, we calculate the difference of the averages within each group. The we use the difference to estimate the distribution mean differences in the two groups.

Why point estimation is important? Learning from samples to make conjecture about population distributions (statistical inference) is one of the basic goals of statistics. Point estimation provides a number to any unknown parameter in the population distribution. All statistical predictions involve estimation of an unknown model that generate the data.

Denote an observed sample of a univariate distribution as $x_1, x_2, \ldots, x_n$, an observed sample of a bivariate distribution as $\{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$ and an observed sample of a multivariate distribution as $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$. Denote a random sample of a univariate distribution as $X_1, X_2, \ldots, X_n$, a random sample of a bivariate distribution as $\{(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)\}$ and a random sample from a multivariate distribution as $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n$. Typically we assume independence among $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n$. The resulting terminology is that $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n$ is a simple random sample.

- **Point estimate** is a single numerical value summarizing the observed sample, for example, a function of $x_1, x_2, \ldots, x_n$.
- **Point estimator** is a random variable that summarizing the random sample, for example, a function of $X_1, X_2, \ldots, X_n$.

We view a point estimate as a realization of its corresponding point estimator. Study properties of the point estimator answers how good the point estimate is.

1.1 Application

Table 1: Commonly seen estimaters. Define $\bar{x} = \sum_{i=1}^n x_i/n$, $s^2 = \frac{1}{n-1}\sum_{i=1}^n (x_i - \bar{x})^2$, $\hat{p}$ as sample proportion of successes, i.e. $\hat{p} = \bar{x}_i$ if $x_i \in \{0, 1\}$, sample correlation $r_{xy} = \frac{\sum_{i=1}^n y_i(x_i - \bar{x})}{\sqrt{\left[\sum_{i=1}^n (y_i - \bar{y})^2 \sum_{i=1}^n (x_i - \bar{x})^2\right]}}$.

	Parameter of interest	Estimate
(1)	The mean μ of a population	$\bar{x}$
(2)	The variance σ^2 of a population	s^2
(3)	Success rate p	$\hat{p}$
(4)	The μ parameter of Normal distribution (μ, σ^2)	$\bar{x}$
(5)	The σ^2 parameter of Normal distribution (μ, σ^2)	s^2
(6)	The λ parameter in Poisson distribution (λT)	$\bar{x}/T$
(7)	The λ parameter in Exponential distribution (λ)	$1/\bar{x}$
(8)	Correlation of two continuous variables	r_{xy}

Example 1.1. For a sample of 20 bottles of wine that are selected from a large batch, wine quality ratings are 19.2, 18.3, 17.1, 15.2, 14.0, 13.8, 12.8, 17.3, 16.3, 16.0, 15.7, 15.3, 14.3, 14.0,13.8, 12.5, 11.5, 14.2,17.3,15.8.

PH level for the 20 bottles are 3.85,3.75,3.88,3.66,3.47,3.75,3.92,3.97,3.76,3.98,3.75,3.77,3.76,3.76,3.90, 3.80,3.65,3.60,3.86,3.93.

- Give an estimate for the mean of wine quality rate of the large batch (μ).
- Give an estimate for the variance of wine quality rate of the large batch (σ^2).
- Give an estimate for the correlation of wine quality and PH.

R codes to find the answers:

```
mean(winequality)
var(winequality)
cor(winequality,PH)
```

Interpretations:

- The average wine quality rate of the batch is estimated to be 15.22.
- The variance of wine quality rate of the batch is estimated to be 3.99.

- The correlation between wine quality rate and PH level is estimated to be 0.349.

Define

- $\bar{X} = \sum_{i=1}^{n} X_i/n$.
- $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$.
- $\hat{P}$ as random sample proportion of successes, i.e. $\hat{P} = \bar{X}_i$ if $X_i \in \{0, 1\}$.
- Random sample correlation $R_{xy} = \frac{\sum_{i=1}^{n} Y_i(X_i - \bar{X})}{\sqrt{\left[\sum_{i=1}^{n}(Y_i-\bar{Y})^2 \sum_{i=1}^{n}(X_i-\bar{X})^2\right]}}$

Table 2: Commonly seen estimates and estimators.

	Parameter of interest	estimate	Estimator
(1)	The mean μ of a population	$\bar{x}$	$\bar{X}$
(2)	The variance σ^2 of a population	s^2	S^2
(3)	Success rate p	$\hat{p}$	$\hat{P}$
(4)	The μ parameter of Normal distribution (μ, σ^2)	$\bar{x}$	$\bar{X}$
(5)	The σ^2 parameter of Normal distribution (μ, σ^2)	s^2	S^2
(6)	The λ parameter in Poisson distribution (λT)	$\bar{x}/T$	$\bar{X}$
(7)	The λ parameter in Exponential distribution (λ)	$1/\bar{x}$	$1/\bar{X}$
(8)	Correlation of two continuous variables	r_{xy}	R_{xy}

Suppose we assume a linear regression model between wine quality Y and PH X:

$$Y_i = \alpha + \beta X_i + \epsilon_i$$

where ϵ_i is assumed to follow a Normal distribution with mean 0 and variance σ^2. How do we estimate α, β, and σ?

- It turns out β can be estimated by $\hat{\beta} = \frac{\sum_{i=1}^{n} y_i(x_i-\bar{x})^2}{\sum_{i=1}^{2}(x_i-\bar{x})^2}$.
- α can be estimated by $\bar{y} - \hat{\beta}\bar{x}$.
- σ^2 can be estimated by $\sum(y_i - \hat{\alpha} - \hat{\beta}x_i)^2/(n-2)$.

R codes to fitting the linear regression model and plot it:

```
plot(PH, winequality)
abline(lm(winequality~PH))
```

1.2 General concepts

Bias of an estimator:the bias of a point estimator $\hat{\theta}$ for parameter θ is $E(\hat{\theta}) - \theta$. If the bias is zero, we say the estimator is unbiased.

Mean squared error is defined as

$$MSE(\hat{\theta}) = E(\hat{\theta} - \theta)^2.$$

We typically select an unbiased estimator with the smallest MSE. Sometimes, we compromise the answer by finding a biased estimator but with the smallest MSE.

Example 1.2. Let $X_1, X_2, \ldots, X_{10}$ be a simple random sample from a Normal distribution $N(2, 3^2)$ and $\bar{X} = \frac{1}{10}\sum_{i=1}^{10} X_i$ is the random sample mean. We want to investigate two estimators for distribution mean μ (pretending we do not know $\mu = 2$): $\bar{X}$ and $\frac{X_1+X_2}{2}$. What are their bias and mean squared errors? Which estimator is preferred?

1.3 Maximum likelihood method

How to find the estimators? Commonly seen methods include maximum likelihood method and moment method. We will focus on the maximum likelihood method and a univariate random sample.

Let x_1, x_2, $\ldots$, x_n be a simple random sample from distribution $f(x)$ where $f(x)$ can be discrete mass function or a continuous probability density function with unknown parameter θ. Then the likelihood function of the parameter is

$$L(\theta) = f(x_1) \times f(x_2) \cdots \times f(x_n).$$

The likelihood function is now a function of only the unknown parameter θ. The maximum

likelihood estimate of θ is value of θ that maximizes the likelihood function $L(\theta)$ or $\log L(\theta)$. For example, if $f(x) = \binom{10}{x} p^x (1-p)^{10-x}$ where p is the unknown parameter, then the likelihood function based on a sample of $\{2, 2, 4, 6\}$ is

$$\begin{aligned} L(p) &= f(x_1) \times f(x_2) \cdots \times f(x_4) \\ &= \left(\binom{10}{2} p^2 (1-p)^8 \right)^2 \binom{10}{4} p^4 (1-p)^6 \binom{10}{6} p^6 (1-p)^4 \\ &= \left(\binom{10}{2} \right)^2 \binom{10}{4} \binom{10}{6} p^{14} (1-p)^{26} \end{aligned}$$

To maximize $L(p)$ with respect to p, we can maximize $log(L(p))$ with respect to p instead. In fact $log(L(p)) = log\left(\left(\binom{10}{2} \right)^2 \binom{10}{4} \binom{10}{6} \right) + 14 * log(p) + 26 * log(1-p)$. Taking derivative of $log(L(p))$ with respect to p, we obtain $\frac{dlog(L(p))}{dp} = 14/p - 26/(1-p)$. Setting the derivative to zero, we solve for $p = 7/20$. Since the second derivative of $log(L(p))$ is less than zero, hence $p = 7/20$ maximizes $log(L(p))$. Hence the maximum likelihood estimate of p is $7/20$.

A maximum likelihood estimator is obtained by maximizing the likelihood with respect to the unknown parameter but we replace $x_1, x_2, \ldots, x_n$ by $X_1, X_2, \ldots, X_n$. Following the previous example,

$$\begin{aligned} L(p) &= f(X_1) \times f(X_2) \times f(X_3) \times f(X_4) \\ &= \binom{10}{X_1} p^{X_1} (1-p)^{10-X_1} \binom{10}{X_2} p^{X_2} (1-p)^{10-X_2} \\ &\times \binom{10}{X_3} p^{X_3} (1-p)^{10-X_3} \binom{10}{X_1} p^{X_4} (1-p)^{10-X_4} \\ &= c p^{X_1+X_2+X_3+X_4} (1-p)^{40-X_1-X_2-X_3-X_4} \end{aligned}$$

Hence $Log(L(p)) = log(c) + \sum_{i=1}^4 X_i log(p) + (40 - \sum_{i=1}^4 X_i) log(1-p)$ and hence $\frac{dLog(L(p))}{dp} = \sum_{i=1}^4 X_i/p - (40 - \sum_{i=1}^4 X_i)/(1-p)$. Setting $\frac{dLog(L(p))}{dp}$ to zero, we solve $p = \sum_{i=1}^4 X_i/40$. The second derivative is less than zero, so setting $p = \sum_{i=1}^4 X_i/40$ maximizes the log likelihood. We call $\sum_{i=1}^4 X_i/40$ an estimator and denote it by $\hat{p}$.

Example 1.3. Consider an exponential distribution with rate parameter λ, i.e. $f_X(x) = \lambda \exp(-\lambda x), x > 0$. Derive the maximum likelihood estimator of λ based on a simple random sample of size n.

2 Chapter 8: Confidence intervals

In statistics, a confidence interval (CI) is a type of interval estimate, computed from the statistics of the observed data, that might contain the true value of an unknown population parameter. The interval has an associated confidence level that, loosely speaking, quantifies the level of confidence that the parameter lies in the interval. More strictly speaking, the confidence level represents the frequency (i.e. the proportion) of possible confidence intervals that contain the true value of the unknown population parameter. In other words, if confidence intervals are constructed using a given confidence level from an infinite number of independent sample statistics, the proportion of those intervals that contain the true value of the parameter will be equal to the confidence level.

A confidence interval provides complementary information addressing the uncertainty of a point estimate. For example, if two samples both gives a point estimate of 0.2 for the distribution mean, one sample is based on 100 observations and its 95% confidence interval is [0.11, 0.31]; the other is based on 10 observations and its 95% confidence interval is [-0.1, 0.4]. Both confidence intervals reflects the uncertainties for the point estimate. Both confidence intervals reflects the uncertainties for the point estimate. We would prefer the first sample result since it is narrower and hence more informative.

Commonly used confidence intervals include:

- Confidence interval on the mean of a normal distribution or a general distribution, assuming variance is known/unknown.
- Confidence interval on the variance of a normal distribution or a general distribution.
- Confidence interval on the difference of two normal distributions or two general distri-

butions, assuming variances known/unknown.

- Confidence interval on regression coefficients.

2.1 Confidence intervals for sample from a Normal distribution

2.1.1 Confidence intervals for the distribution mean

Conditions for the z-confidence interval formula:

(a) We can regard our data as a simple random sample from the population, i.e. the sample units were selected with equal probability.

(b) The distribution of the variable we measure has an exactly Normal distribution $N(\mu, \sigma^2)$ in the population.

(c) We don't know the population mean μ. But we do know the population variance σ^2. Knowing the population variance is rare, but unless the sample size is small, using an approximate answer is also okay.

Let q be a number between 0 and 1, and z_q be a number denotes the $100\times(1-q)$ percentile of the standard normal distribution. For example, $z_{0.025}$ denote the 97.5% percentile, which is approximately 1.96. Denote $\bar{x}$ as the sample mean of an observed sample $(x_1, \ldots, x_n)$. A $100(1-\alpha)\%$ CI on μ is given by

$$[\bar{x} - z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}, \bar{x} + z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}]$$

For example, if a sample is $(1, 3, -1, 2, 5)$ and assume $\sigma^2 = 4$, then $n = 5$, $\bar{x} = 2.4$, the 95% confidence interval for the Normal distribution mean is

$$(2.4 - 1.96 * 2/\sqrt{5}, 2.4 + 1.96 * 2/\sqrt{5}) = (0.65, 4.15).$$

Interpretation: the confidence interval $0.65, 4.15]$ contains the true value of μ (interpret μ in the context, for example, mean income level), with 95% confidence.

Addressing the assumptions

- Can we view our sample data as a simple random sample? Typically taking a sample randomly from a very large population can be viewed as a simple random sample.

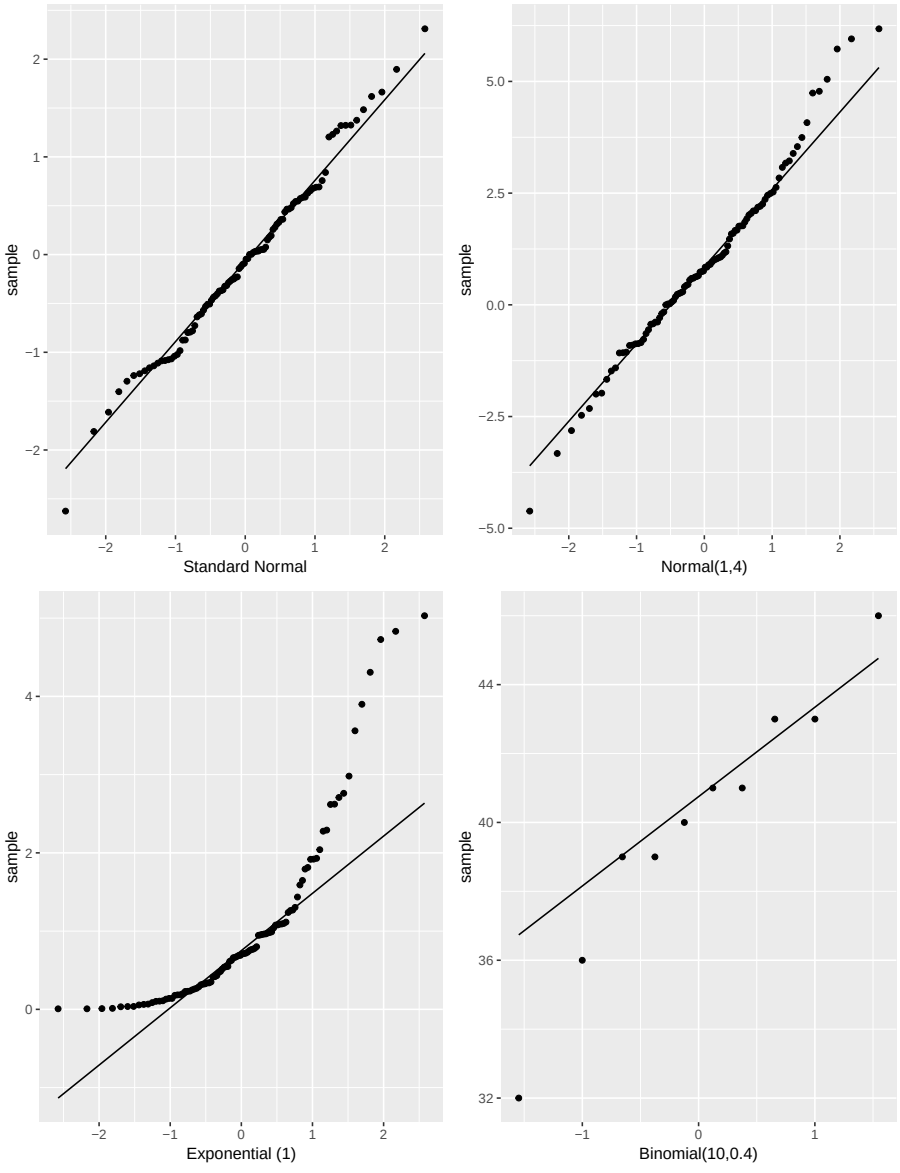
- Is the distribution of the variable we measure Normally distributed? We can use QQ plot and a formal test.

A QQ plot is a scatterplot of the sorted sample $x_{(i)}$ for $i = 1, \ldots, n$ against standard Normal distribution percentiles for 100*$p_i = \frac{i}{n} - \frac{1}{2n}$. Use library (ggplot2) and then command

```
ggplot()+stat_qq(aes(sample = x1))+stat_qq_line(aes(sample = x1))
```

For a large sample from standard Normal distribution the plot should be a straight line through the origin with slope 1. If the plot is a straight line with a different slope or intercept, then the data distribution corresponds to a general Normal distribution.

Figure 1: QQ plot of four simulated datasets where each dataset has 100 data points.



Example 2.1. The Charpy V-notch (CVN) technique measures impact energy and is often used to determine whether or not a material experience a ductile-to-brittle transition with decreasing temperature. Ten measurements of impact energy (J) on specimens of a 238 Steel cut at 60^o are as follows: 64.1, 64.7, 64.5, 64.6, 64.5, 64.3, 64.6, 64.8, 64.2, 64.3. Assume that impact energy is normally distributed with $\sigma = 1J$. Based on the data $\bar{x} = 64.46$, $n = 10$. Find the 95% CI for μ and assess the Normality assumption.

What if population is unknown? We can use a t-confidence interval. Suppose we observe sample $(x_1, \ldots, x_n)$ and $\bar{x}$ is the sample mean and s is the sample standard deviation. The $100(1-\alpha)\%$ t-confidence interval of μ is

$$[\bar{x} - t_{\frac{\alpha}{2},n-1}\frac{s}{\sqrt{n}}, \bar{x} + t_{\frac{\alpha}{2},n-1}\frac{s}{\sqrt{n}}]$$

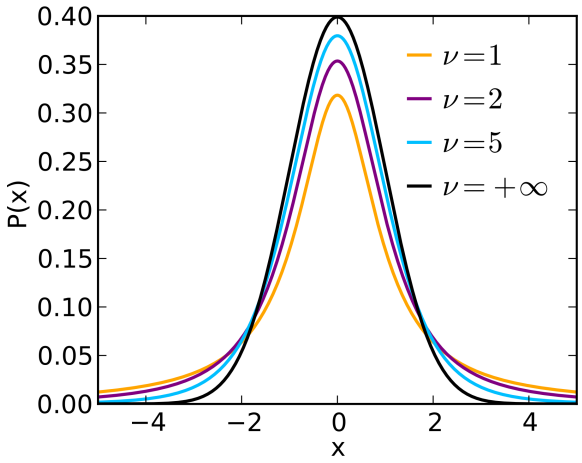
where $t_{\frac{\alpha}{2},n-1}$ is the $100 \times (1 - \frac{\alpha}{2})$ percentage point of a t distribution with $n-1$ degrees of freedom. In R, use "qt($\frac{\alpha}{2}$, $n-1$, lower.tail = FALSE, log.p = FALSE)" to obtain the value. Hence for the mean impact energy example, the t-confidence interval is

$$[64.46 - 2.26 * 0.227/\sqrt{10}, 64.46 + 2.26 * 0.227/\sqrt{10}] = [64.30, 64.62]$$

What is t-distribution? It is also called "student's t-distribution". A t distribution with ν degree of freedom has PDF

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

Figure 2: t-distributions.



2.1.2 Method

The confidence interval is based on the sampling distribution of $\bar{X}$: if $X_1, X_2, \ldots, X_n$ is a simple random sample from Normal distribution with mean μ and variance σ^2, i.e. $X_i \sim N(\mu, \sigma^2), i = 1, \ldots, n$ and they are mutually independent, then for any random sample size n,

$$\bar{X} \sim N(\mu, \sigma^2/n),$$

or equivalently,

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

where $\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n}$

Therefore,

$$P(-z_{\frac{\alpha}{2}} \leq Z \leq z_{\frac{\alpha}{2}}) = 1 - \alpha. \tag{1}$$

Plug $Z = \frac{\bar{X}-\mu}{\sigma/\sqrt{n}}$ into formula (1), we get

$$P\left(-z_{\frac{\alpha}{2}} \leq \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \leq z_{\frac{\alpha}{2}}\right) = 1 - \alpha \text{ and then}$$
$$P\left(\bar{X} - z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha.$$

Let $L = \bar{X} - z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}$ and $U = \bar{X} + z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}$, then $P(L \leq \mu \leq U) = 1 - \alpha$. It means that the random intervals will have probability $100(1-\alpha)\%$ to cover the truth.

The confidence interval is based on the sampling distribution of $\frac{\bar{X}-\mu}{S/\sqrt{n}}$: if $X_1, X_2, \ldots, X_n$ is a random sample from Normal distribution with mean μ and variance σ^2, i.e. $X_i \sim N(\mu, \sigma^2), i = 1, \ldots, n$ and they are mutually independent, then for any random sample size n,

$$T = \frac{\bar{X} - \mu}{S/\sqrt{n}} \sim t_{n-1},$$

where $\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n}$, $S = \sqrt{\frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1}}$, and t_{n-1} represents a t distribution with $n-1$ degrees of freedom.

$$P\left(-t_{\frac{\alpha}{2},n-1} \leq \frac{\bar{X} - \mu}{S/\sqrt{n}} \leq t_{\frac{\alpha}{2},n-1}\right) = 1 - \alpha \text{ and then}$$
$$P\left(\bar{X} - t_{\frac{\alpha}{2},n-1}\frac{S}{\sqrt{n}} \leq \mu \leq \bar{X} + t_{\frac{\alpha}{2},n-1}\frac{S}{\sqrt{n}}\right) = 1 - \alpha.$$

Let $L = \bar{X} - t_{\frac{\alpha}{2},n-1}\frac{S}{\sqrt{n}}$ and $U = \bar{X} + t_{\frac{\alpha}{2},n-1}\frac{S}{\sqrt{n}}$, then $P(L \leq \mu \leq U) = 1 - \alpha$. It means that the random intervals will have probability $100(1-\alpha)\%$ to cover the true μ.

2.1.3 Sample size calculation

Suppose we are still interested in to know the distribution mean μ. We want to plan ahead before data are collected. One question is: how many subjects do we include in the study? First we start with a pre-specified amount of error that we allow in estimating μ. Denote the error by E. Note that $\bar{X}$ is an unbiased estimator of μ. The variance of $\bar{X}$ is σ^2/n. Hence, though we expect $\bar{x}$ to be right at the target, but because of sampling variability, a one time estimate is usually in the neighborhood of μ, and the size of the neighborhood depend on n and σ^2. We want to calculate the smallest sample size so that $|\bar{x} - \mu| <= E$ with $100(1-\alpha)\%$ confidence.

Assume σ^2 is known (usually this is unknown, but an estimate of it can be used). The sample size needed to ensure the **absolute error** $|\bar{x} - \mu|$ will not exceed a specified amount E with $100(1-\alpha)\%$ confidence is

$$n \geq \left(\frac{z_{\frac{\alpha}{2}}\sigma}{E}\right)^2.$$

How to derive this formula? From confidence interval formula, we know $|\bar{x} - \mu| \leq z_{\frac{\alpha}{2}}\sigma/\sqrt{n}$

with $100(1-\alpha)\%$ confidence. Hence let $z_{\frac{\alpha}{2}}\sigma/\sqrt{n} \leq E$ and solve for n, we will obtain $n \geq \left(\frac{z_{\frac{\alpha}{2}}\sigma}{E}\right)^2$.

Example 2.2. Consider the CVN test again and suppose that we want to determine how many specimens must be tested to ensure that with 95% that the absolute error does not exceed 0.5, i.e. $|\bar{x}-\mu| \leq 0.5$, i.e. the smallest sample size to ensure that with 95% that the absolute error does not exceed 0.05. Use the sample standard deviation in replace of σ.

2.1.4 Confidence intervals for the distribution variance

Confidence interval for the variance is useful in uncertainty quantification. In experiment data, we compare the variability due to different sources. Previously we used sample variance to estimate variance. Confidence interval for the variance acknowledges the sampling variability in estimating variance. Suppose data $x_1, x_2, \ldots, x_n$ are simple random samples from a Normal distribution. The $100(1-\alpha)\%$ for σ^2 is

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

where $\chi^2_{n-1,1-\alpha/2}$ is the $1-\alpha/2$ percentile of a Chi-square distribution that has a $n-1$ degree of freedom; similarly, $\chi^2_{n-1,\alpha/2}$ is the $\alpha/2$ percentile of a Chi-square distribution that has a $n-1$ degree of freedom. The $100(1-\alpha)\%$ for σ is

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

Example 2.3. The Charpy V-notch (CVN) technique measures impact energy and is often used to determine whether or not a material experience a ductile-to-brittle transition with decreasing temperature. Ten measurements of impact energy (J) on specimens of a 238 Steel cut at 60^o are as follows: 64.1, 64.7, 64.5, 64.6, 64.5, 64.3, 64.6, 64.8, 64.2, 64.3. Obtain a

95% confidence interval for σ^2.

2.2 Large-sample confidence interval on the mean

In this section, we present a large-sample CI for population mean μ that does not require Normal assumption.

Simple conditions for inference about a mean in this section:

(a) We can regard our data as a simple random sample from the population.

(b) Generally, sample size n should be at least 40 to use this result reliably.

(c) Both mean μ and standard deviation σ could be unknown.

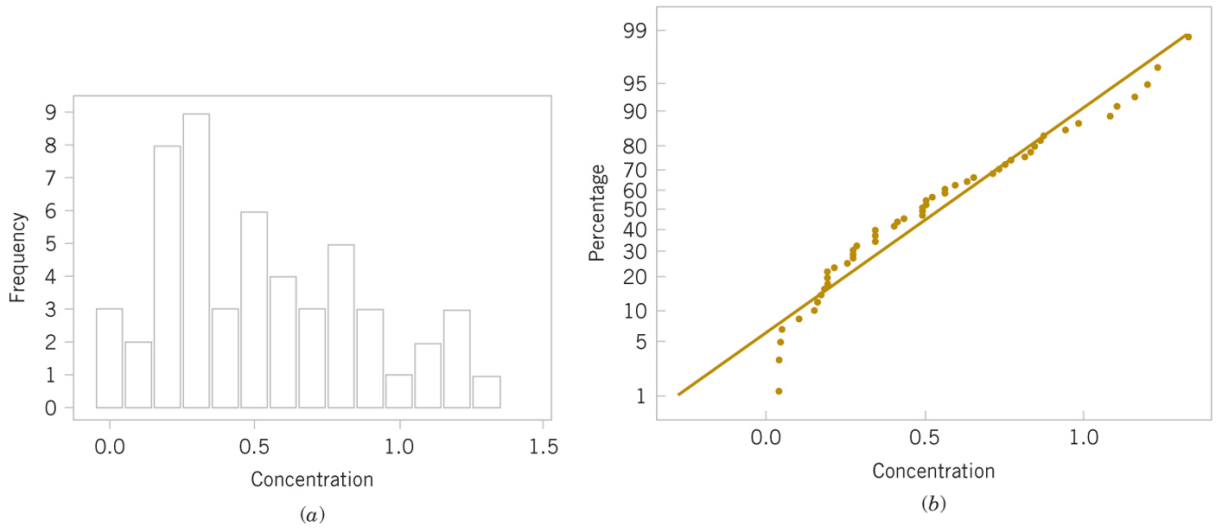
A large-sample confidence interval for μ with confidence level of approximately $100(1-\alpha)\%$ is

$$[\bar{x} - z_{\frac{\alpha}{2}} \frac{s}{\sqrt{n}}, \bar{x} + z_{\frac{\alpha}{2}} \frac{s}{\sqrt{n}}].$$

The confidence interval is based on **large sample** sampling distribution of $\frac{\bar{X}-\mu}{S/\sqrt{n}}$: when n is large (usually > 30), the quantity $\frac{\bar{X}-\mu}{S/\sqrt{n}}$ has an approximate standard Normal distribution. The sample size here is just a rule of thumb. Better confidence interval formula is based on Bootstrapping, which does not rely on this assumption.

Example 2.4. An article in the 1993 volume of the transactions of the American Fisheries Society reports the results of a study to investigate the mercury contamination in large mouth bass. A sample of fish was elected from 53 Florida lakes, and mercury concentration in the muscle tissue was measured (ppm). Find an approximate 95% CI on μ. The summary statistics for the sample are as follows:

Variable	N	Mean	StDev
Concentration	53	0.525	0.349



2.3 Large sample confidence interval of a population proportion

Population proportion refers to is a parameter that describes a percentage value associated with a population. For example, the 2010 United States Census showed that 83.7% of the American Population was identified as not being Hispanic or Latino. The value of .837 is a population proportion. In general, the population proportion and other population parameters are unknown. A census can be conducted in order to determine the actual value of a population parameter, but often a census is not practical due to its costs and time consumption.

A population proportion is usually estimated through a simple random sample proportion obtained from an observational study or experiment. For example, the National Techno-

logical Literacy Conference conducted a national survey of 2,000 adults to determine the percentage of adults who are economically illiterate. The study showed that 72% of the 2,000 adults sampled did not understand what a gross domestic product is.The value of 72% is a sample proportion. A population proportion can also be estimated through the usage of a confidence interval known as a one-sample proportion Z-interval.

Denote the random sample as $x_1, x_2, \ldots, x_n$, each taking value 0 or 1. The theoretical framework is to view each sample item x_i which takes value 0 or 1 as a realization from a Bernoulli random variable. Consider $X_1, X_2, \ldots, X_n$ are a random sample from Bernoulli population. Suppose the proportion of "1" in the population is p, then distribution of each X_i is Bernoulli (p). The Bernoulli distribution has mean p and variance $p(1-p)$. Consider $\bar{X} = \frac{\sum_{i=1}^{p} X_i}{n}$, then according to the central limit theorem,

$$\bar{X} \sim N\left(p, \frac{p(1-p)}{n}\right)$$

Denote $\hat{p}$ as the sample proportion. Denote $\hat{p}$ as the sample proportion. It is calculated as $\hat{p} = \frac{\sum_{i=1}^{n} x_i}{n}$. Apply the large sample confidence interval formula, we obtain a one-sample proportion Z-interval for p with confidence level of $100(1-\alpha)\%$

$$\left[\hat{p} - z_{\frac{\alpha}{2}}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}, \hat{p} + z_{\frac{\alpha}{2}}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right],$$

where $z_{\frac{\alpha}{2}}$ is the $100(1-\frac{\alpha}{2})$ percentile of standard Normal disribution.

Use this interval only when the numbers of successes and failures in the sample are both at least 5 (this is just a rule of thumb, more accurate confidence interval can be obtained through Bootstrap method.).

Example 2.5. In a random sample of 85 automobile engine crankshaft bearings, 10 have a surface finish that is rougher than the specifications allow. What would be the point estimate of the proportion of bearings in the population that exceeds the roughness specification? Find a 95% confidence interval for the proportion.

If we desire a bound of error $|p - \hat{p}|$, i.e. $|p - \hat{p}| < E$, the sample size needed is

$$n \geq \left(\frac{z_{\frac{\alpha}{2}}}{E}\right)^2 p(1-p).$$

Usually, we plug in p using values from pe-studies or fix p at 0.5 to obtain the largest sample size needed.

In the automobile engine example, what is the minimum sample size needed if we want to be 95% confident that the error in using $\hat{p}$ to estimate p is less than 0.05? Using the number from the random sample, $n \geq \left(\frac{1.96}{0.05}\right)^2 0.118(1 - 0.118) = 159.9$. The smallest sample size is 160. To be more conservative, $n \geq \left(\frac{1.96}{0.05}\right)^2 0.118(1 - 0.118) = 284.16$. The smallest sample size is 285.