Statistics for Engineers Lecture 6
One-Sample Inference

Chong Ma

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

March 20, 2017

Outline

One Sample inference

In this lecture, we discuss one-sample inference procedures for three population parameters:

- A population **mean** μ.
- A population **variance** σ^2.
- A population **proportion** p.

Note that these are population-level quantities, so they are unknown. Our goal is to use sample information to estimate these quantities. To begin with, let us consider estimate a **population mean** μ. In the last lecture, we know that $\bar{Y}$ is an **unbiased estimator** for μ, whatever the population distribution is. However, reporting $\bar{Y}$ alone does not acknowledge that there is a variability attached to the estimator.

One Sample Inference

Recall the pipe example in lecture 5, with $n = 25$ measured pipes, reporting $\bar{y} \approx 1.299$ as an estimate of the population mean μ does not account for the fact that

- the 25 pipes measured were drawn randomly from a population of all pipes.
- different sample would give different sets of pipes(leading to different values of $\bar{y}$).

To address this problem, we therefore pursue the topic of **interval estimation**(also known as **confidence interval**). The main difference between a point estimate(like $\bar{y} \approx 1.299$) and an interval estimate is that

- a **point estimate** is a "one-shot guess" for the parameter, ignoring the variability in the estimate.
- an **interval estimate** is an interval of values, by taking the point estimate and then adjusting it downwards and upwards to account for the point estimate's variability.

Outline

CI for population mean μ

Recall that if $Y_1, Y_2, \ldots, Y_n \overset{i.i.d}{\sim} N(\mu, \sigma^2)$, then the quantity

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

We use this sampling distribution of $\bar{Y}$ to create an interval estimate for the population mean μ. A $100(1-\alpha)$ percent confidence interval for the population mean μ is

$$[\bar{Y} \pm t_{n-1,\alpha/2}\frac{S}{\sqrt{n}}] = [\bar{Y} - t_{n-1,\alpha/2}\frac{S}{\sqrt{n}}, \bar{Y} + t_{n-1,\alpha/2}\frac{S}{\sqrt{n}}]$$

Where $t_{n-1,\alpha/2}$ is the **upper** $\alpha/2$ quantile from t_{n-1} pdf. That is, $100(1-\alpha/2)$ percent of values from t_{n-1} pdf fall below or at $t_{n-1,\alpha/2}$. Because t_{n-1} pdf is symmetric about zero, $t_{n-1,1-\alpha/2} = -t_{n-1,\alpha/2}$.

CI for population mean μ

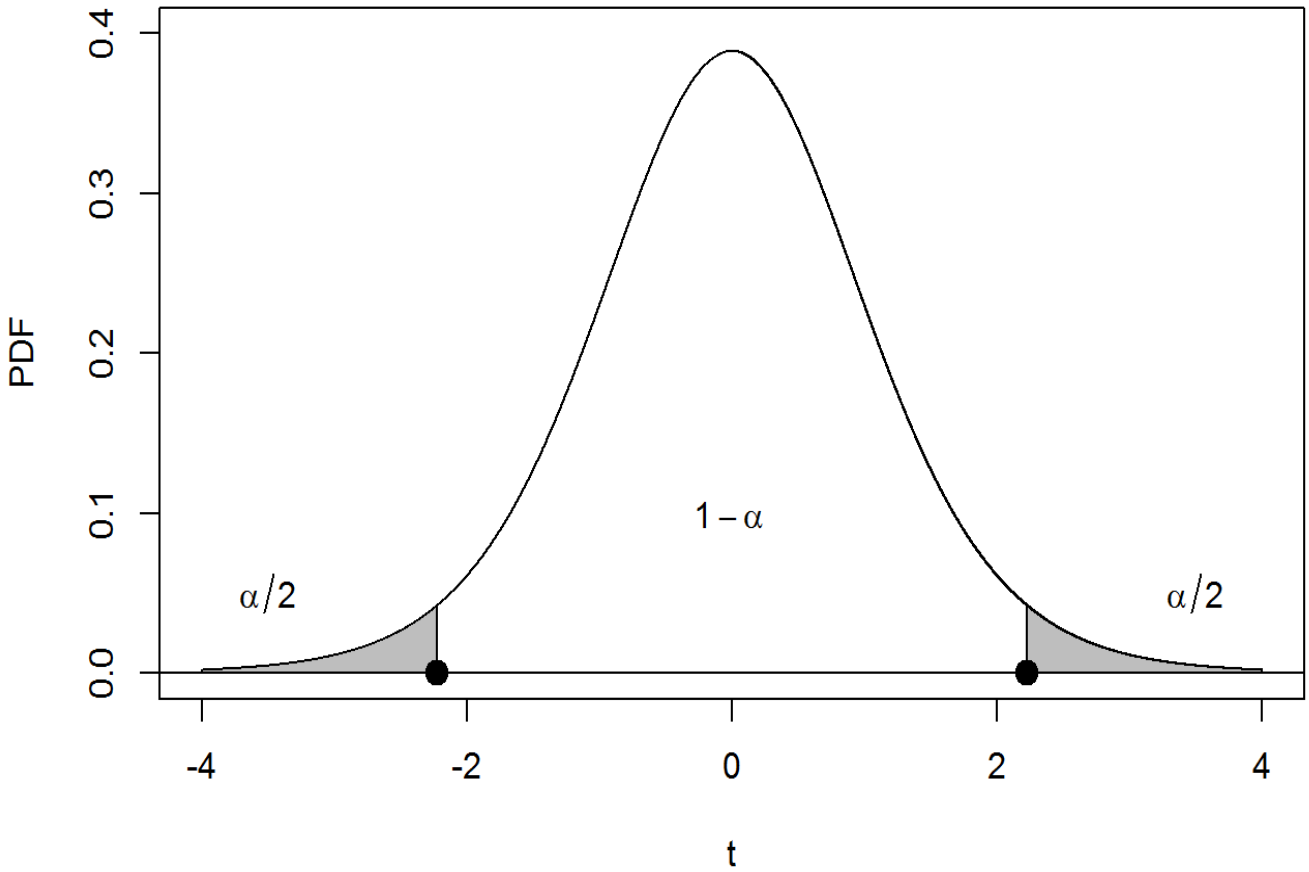

Figure 1: A t pdf with $n-1$ degrees of freedom. The upper $\alpha/2$ and lower $\alpha/2$ areas were shaded.

CI for population mean μ

- The form of the interval:

$$\underbrace{\text{point estimate}}_{\bar{Y}} \pm \underbrace{\text{quantile}}_{t_{n-1,\alpha/2}} \times \underbrace{\text{standard error}}_{S/\sqrt{n}}$$

- Here is how we interpret this interval: We say **"We are** $100(1-\alpha)$ **percent confident that the population mean** μ **is in this interval**.
- The word "confident" does not mean "probability". The term **"confidence"** means that **if we were able to sample from the population over and over again, each time computing a** $100(1-\alpha)$ **percent confidence interval for** μ**, then** $100(1-\alpha)$ **percent of the intervals we would compute would contain the population mean** μ.
- In other words, **"confidence"** refers to **"long term behavior"** of many intervals; not probability for a single interval. Because of this, we call $100(1-\alpha)$ the **confidence level**.

CI for population mean μ

- Typical confidence levels are
  - 90 percent($\alpha = 0.10$)
  - 95 percent($\alpha = 0.05$)
  - 99 percent($\alpha = 0.01$)
- The **length** of the $100(1-\alpha)$ percent confidence interval $[\bar{Y} \pm t_{n-1,\alpha/2}\frac{S}{\sqrt{n}}]$ is equal to $2t_{n-1,\alpha/2}\frac{S}{\sqrt{n}}$. Hence, conditioning on other things,
  - The larger the sample size n, the smaller the interval length.
  - The smaller the population variance σ^2, the smaller the interval length.
  - The larger the confidence level $100(1-\alpha)$, the larger the interval length.

Remark: Clearly, shorter confidence intervals are preferred, because they are more informative. Although lower confidence levels will produce shorter intervals, it pays a price in that you have less confidence that your interval contains μ.

CI for population mean μ

Example Acute exposure to cadmium produces respiratory distress and kidney and liver damage(and possible death). For this reason, the level of airborne cadmium dust and cadmium oxide fume in the air, denoted by Y (measured in milligrams of cadmium per m^3 of air), is closely monitored. A random sample of $n = 35$ measurements from a large factory are given as follows (see completed data in R tutorial).

0.044	0.030	0.052	0.044	0.046	0.020	0.066
...	...	...	...	...	...	...
0.053	0.060	0.047	0.051	0.054	0.042	0.051

Based on the data above, find a 99 percent confidence interval for μ, the population mean level of airborne cadmium.
Using R, we can readily compute the sample mean and sample standard deviation, which are

$$\bar{y} = 0.049, s = 0.011$$

CI for population mean μ

For $\alpha = 0.01$, the according quantile is $t_{34,0.01/2} \approx 2.728$(using R code **qt(1-0.01/2,34)**). Applying the interval formula $[\bar{Y} \pm t_{n-1,\alpha/2}\frac{S}{\sqrt{n}}]$, we get

$$[0.049 \pm 2.728(\frac{0.011}{\sqrt{35}})] \Rightarrow (0.044, 0.054)\text{mg/m}^3$$

Interpretation: We are 99 percent confident that the population mean level of airborne cadmium μ is between 0.044 and 0.054 mg/m^3.
Remarks: The confidence interval $[\bar{Y} \pm t_{n-1,\alpha/2}\frac{S}{\sqrt{n}}]$ requires two assumptions:

1. $Y_1, Y_2, \ldots, Y_n$ is a random sample.
2. The population distribution is $N(\mu, \sigma^2)$

Recall from the last lecture that the t sampling distribution does still hold approximately even if the underlying population distribution is not perfectly normal. Therefore, the confidence interval(which was derived from this sampling distribution) is also "robust to normality departures".

CI for population mean μ

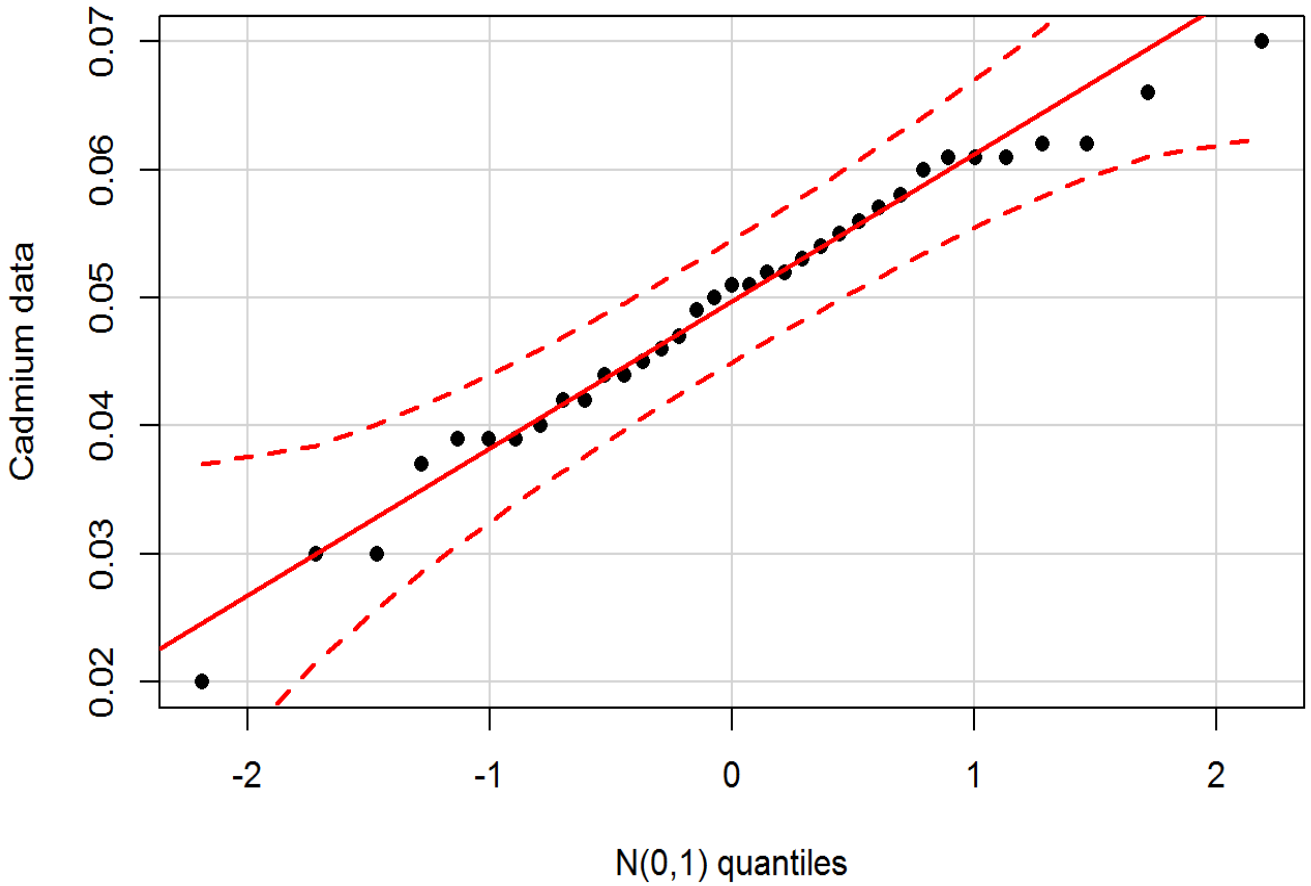

Figure 2: Normal QQ plot for the cadmium data does not reveal any serious departures from the normality assumption.

Outline

CI for population variance σ^2

In many situations, we are concerned not with the mean of a population, but with the variance σ^2 instead. If the population variance σ^2 is excessively large, this could imply a potential problem with a manufacturing process, e.g., where there is too much variability in the measurements produced.

- In a laboratory setting, engineers might wish to estimate the variance σ^2 attached to a measurement system(like scale, caliper, etc).
- In field trials, agronomists are often interested in comparing the variability levels for different cultivars or genetically-altered varieties.
- In clinical trials, physicians are often concerned if there are substantial in the variation levels of patient responses at different clinic sites.

CI for population variance σ^2

If $Y_1, Y_2, \ldots, Y_n \overset{i.i.d}{\sim} \mathcal{N}(\mu, \sigma^2)$, then the quantity

$$Q = \frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1} (= \text{Gamma}(\alpha = \frac{n-1}{2}, \lambda = \frac{1}{2}))$$

is a χ^2 distribution with n-1 degrees of freedom. The χ^2 pdf has the following characteristics:

- It is continuous, skewed to the right and always positive.
- It is indexed by a value of ν called the **degrees of freedom**. In practive, ν is often an integer(related to sample size).
- The χ^2_ν is essentially the Gamma distribution with the shape parameter $\alpha = \frac{\nu}{2}$ and the rate parameter $\lambda = \frac{1}{2}$.
- In R, use **pchisq**(x, ν) to compute the CDF $F(x)$ and use **qchisq**(p, ν) to compute the $100 \times (p)$ percentile for χ^2_ν.

CI for population variance σ^2

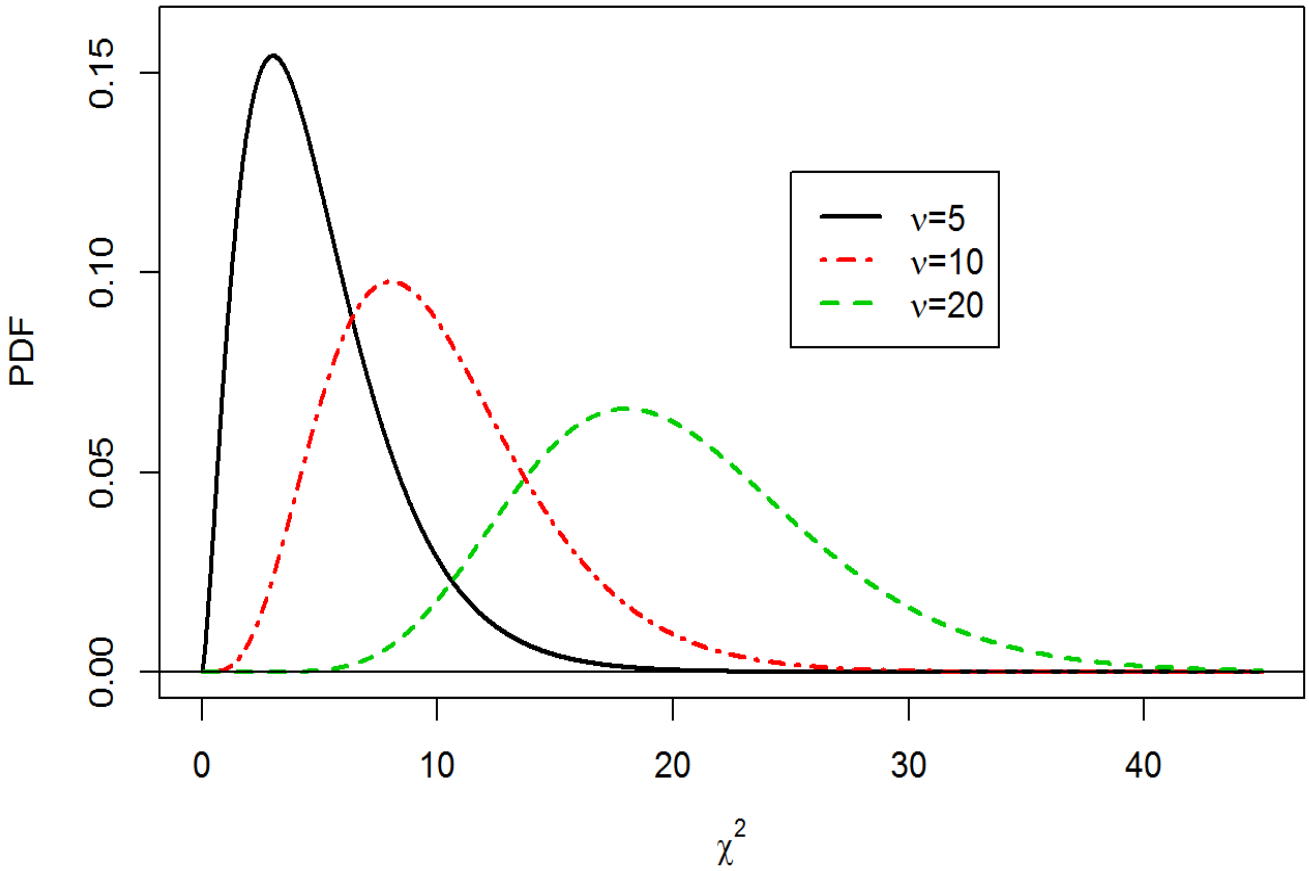

Figure 3: χ^2_ν pdfs with different degrees of freedom.

CI for population variance σ^2

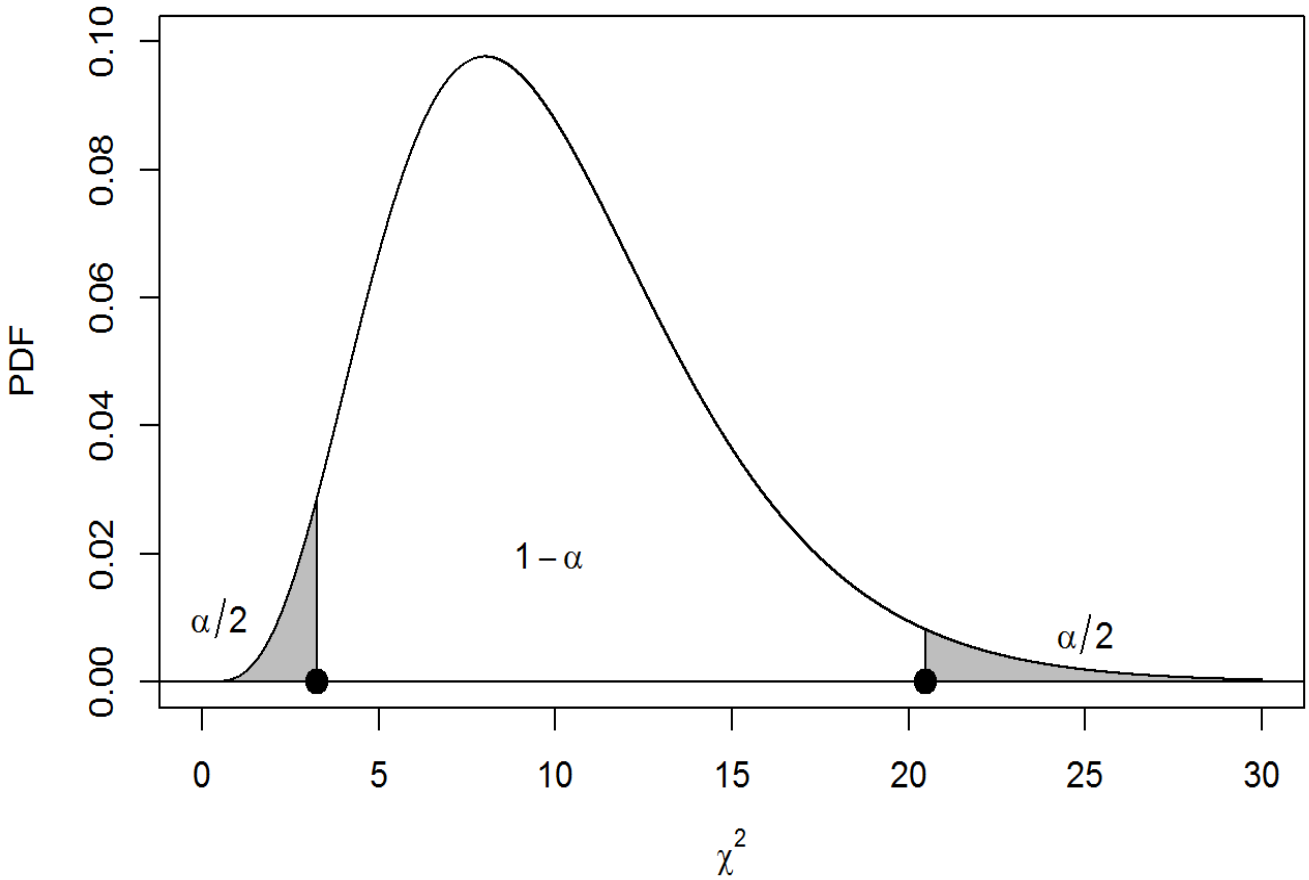

Figure 4: A χ^2 pdf with $n-1$ degrees of freedom. The upper $\alpha/2$ and lower $\alpha/2$ areas are shaded.

CI for population variance σ^2

The $100(1-\alpha)$ percent confidence interval for the population variance σ^2 is

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

We interpret the interval in the same way: **"We are 100$(1-\alpha)$ percent confident that the population variance σ^2 is in this interval."** Note that a $100(1-\alpha)$ percent confidence interval for the **population standard deviation** σ arises from simply taking square root of the endpoints of the σ^2 interval, that is,

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

Remarks: In practice, the interval may be preferred over the σ interval, because standard deviation is a measure of variability in terms of the original units(e.g., dollars, inches, days, etc).

CI for population variance σ^2

Example. Industrial engineers at IKEA observed a random sample of $n = 36$ rivet-head screws used in the Billy Bookcase system. The observed diameters of the top of the screws (measured in cm) are given as following(See completed data in R tutorial):

1.206	1.190	1.200	1.195	1.201
...	...	...	...	...
1.204	1.202	1.196	1.211	1.204

The IKEA manufactured specifications dictate that the population standard deviation diameter for these screws should be **no larger than** $\sigma = 0.003$. Otherwise, there is too much variability in the screws (which could lead to difficulty in construction and hence customer dissatification). Based on the data above, find a 95 percent confidence interval for the population standard deviation σ.

CI for population variance σ^2

Note that $n = 36$, $s^2 = 2.35 \times 10^{-5}$, $\chi^2_{35,1-0.05/2} = 53.20$ and $\chi^2_{35,0.05/2} = 20.57$. Applying the confidence interval formula for the population variance σ^2, we get

$$\left(\sqrt{\frac{(n-1)S^2}{\chi^2_{n-1,1-\alpha/2}}}, \sqrt{\frac{(n-1)S^2}{\chi^2_{n-1,\alpha/2}}}\right) = \left(\sqrt{\frac{35(2.35 \times 10^{-5})}{53.20}}, \sqrt{\frac{35(2.35 \times 10^{-5})}{20.57}}\right)$$

$$= (0.0039, 0.0063)$$

Interpretation: We are 95 percent confident that the population standard deviation σ for the screw diameters is between 0.0039 and 0.0063 cm. This interval suggests that the population standard deviation is larger than 0.003cm, which indicates that there is excessive variability in the diameters of the screws.

CI for population variance σ^2

Remarks: The confidence interval

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

for the population variance σ^2 was created based on the following assumptions:

1. $Y_1, Y_2, \ldots, Y_n$ is a random sample.
2. The population distribution is $N(\mu, \sigma^2)$

Unlike the t confidence interval for a population mean μ, the χ^2 interval for a population variance σ^2 is **not robust** to normally departures. This is true because the sampling distribution

$$Q = \frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}$$

depends critically on the $\mathcal{N}(\mu, \sigma^2)$ population distribution assumption.

CI for population variance σ^2

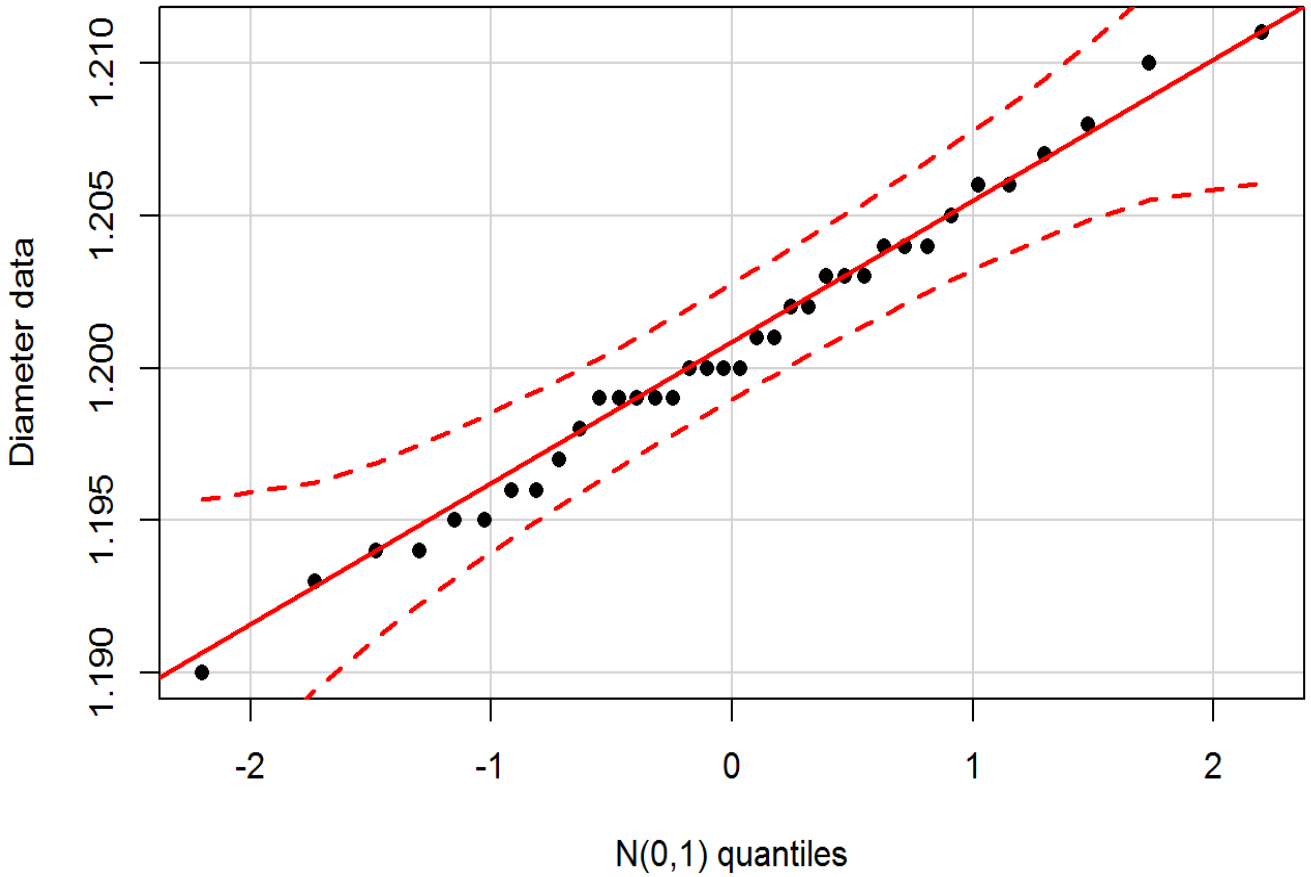

Figure 5: The normal qq plot for IKEA does not indicate serious departures from normality assumption.

Outline

CI for population proportion p

We now switch gears and focus on a new population-level parameter: the **population proportion** p. This parameter is relevant when the characteristic we measure on each individual is **binary**. For example,

- $p =$ proportion of defective circuit boards
- $p =$ proportion of customers who are "satisfied"
- $p =$ proportion of payments received on time
- $p =$ proportion of HIV positives in SC

Recall the **Bernoulli trial** assumptions for each individual in the sample:

1. Each individual results in a "success" or "failure".
2. The individuals are independent.
3. The probability of "success" p is the same for every individual.

For the examples above, "success" are circuit board defective, customer satisfied, payment received on time and HIV positive individual.

CI for population proportion p

Recall that if the individual success/failure statuses in the sample adhere to the Bernoulli trial assumptions, then

$$Y = \text{the number of successes out of n sampled individuals}$$

follows the binomial distribution, i.e., $Y \sim \text{Binomial}(n, p)$. The statistical problem at hand is to use the information in Y to **estimate** p. A natural point estimator for p(**population proportion**) is

$$\hat{p} = \frac{Y}{n}$$

the **sample proportion**. This statistic is simply the proportion of "successes" in the sample(out of n individuals). Note that we have the following mathematical results:

$$E(\hat{p}) = p, \quad se(\hat{p}) = \sqrt{\frac{p(1-p)}{n}}$$

CI for population proportion p

Recall the Central Limit Theorem(CLT), we can say that

$$\hat{p} \sim \mathcal{AN}(p, \frac{p(1-p)}{n})$$

when the sample size n is large. Standardizing $\hat{p}$, we get

$$Z = \frac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{n}}} \sim \mathcal{AN}(0, 1)$$

an approximate **standard normal distribution**. The $100(1-\alpha)$ **percent confidence interval** for the population proportion p is constructed by

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

CI for population proportion p

- Because of p unknown, we replace the p with $\hat{p}$ in $se(\hat{p}) = \sqrt{\frac{p(1-p)}{n}}$ in order to get the confidence interval for the population proportion p.
- The form of the interval is again

$$\underbrace{\text{point estimate}}_{\hat{p}} \pm \underbrace{\text{quantile}}_{z_{\alpha/2}} \times \underbrace{\text{standard error}}_{\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}}$$

- We interpret the interval in the same way: "We are $100(1-\alpha)$ percent confident that the population proportion p is in this interval."
- A common **rule of thumb** for "n is large" is to require

$$n\hat{p} \geq 5, n(1-\hat{p}) \geq 5$$

Under these conditions, the CLT should adequately describe the sampling distribution of $\hat{p}$, thereby making the confidence interval formula above approximately valid.

CI for population proportion p

Example. One source of water pollution is gasoline leakage from underground storage tanks. In Pennsylvania, a random sample of $n = 74$ gasoline stations is selected from the state and the tanks are inspected; 10 stations are found to have at least one leaking tank. Calculate a 95 percent confidence interval for p, the population proportion of gasoline stations with at least one leaking tank.

CI for population proportion p

Example. One source of water pollution is gasoline leakage from underground storage tanks. In Pennsylvania, a random sample of $n = 74$ gasoline stations is selected from the state and the tanks are inspected; 10 stations are found to have at least one leaking tank. Calculate a 95 percent confidence interval for p, the population proportion of gasoline stations with at least one leaking tank.

Solution: In this example, we interpret that

- individual "trial" = gasoline station
- "success" = at least one leaking tank
- p = population proportion of stations with at least one leaking tank

Note that $n = 74, \hat{p} = \frac{10}{74} \approx 0.135, z_{0.05/2} \approx 1.96$, therefore the 95 percent confidence interval for p is

CI for population proportion p

$$\left[\hat{p} \pm z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right] \Rightarrow \left[0.135 \pm 1.96\sqrt{\frac{0.135(1-0.135)}{74}}\right]$$
$$\Rightarrow (0.057, 0.213)$$

Interpretation: We are 95 percent confident that the population proportion of stations in Pennsylvania with at least one leaking tank is between 0.06 and 0.21.

CLT approximation check: We have

$$n\hat{p} = 74(\frac{10}{74}) = 10 \geq 5$$
$$n(1-\hat{p}) = 74(1-\frac{10}{74}) = 64 \geq 5$$

Both of number of successes and failures in the sample are larger than 5 $\Rightarrow$ We can feel decent in using this confidence interval formula.

Outline

Sample size determination

Motivation: In the planning stages of an experiment or investigation, we need to first determine **how many individuals** are needed to compute a confidence interval with a given level of precision. For example, we might want to construct a

- 95 percent confidence interval to estimate the population mean time needed for patients to recover from infection. How many patients should we recruit?
- 99 percent confidence interval to estimate the population proportion of defective parts. How many parts should be sampled?

Note that collecting data almost always costs money. One must be cognizant not only of the statistical issues associated with **sample size determination**, but also of the practical issues like cost, time spent in data collection, personnel training, etc.

Sample size determination about μ

Setting: Suppose that $Y_1, Y_2, \ldots, Y_n \overset{i.i.d}{\sim} \mathcal{N}(\mu, \sigma^2)$. We derived a $100(1-\alpha)$ percent confidence interval for μ, that is,

$$\bar{Y} \pm t_{n-1,\alpha/2}\frac{S}{\sqrt{n}}$$

Suppose that σ^2 is known(albeit it is barely the case in real life), then a $100(1-\alpha)$ percent confidence interval for μ can be calculated by

$$\bar{Y} \pm z_{\alpha/2}\frac{S}{\sqrt{n}}$$

Where denote $B = z_{\alpha/2}\frac{S}{\sqrt{n}}$ by **margin of error**, which is a quantity of variability for $\bar{Y}$ used to estimate μ under the constraint that the probability of such an interval is $1-\alpha$.

Sample size determination about μ

In the above setting, it is possible for us to determine the sample size n necessary once we specify these three pieces of information.

- the value of σ^2(e.g., an educated guess or from historical data)
- the confidence level $100(1-\alpha)$
- the margin of error, B

It is true because

$$B = z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \Leftrightarrow n = \left(\frac{z_{\alpha/2}\sigma}{B}\right)^2$$

This is the necessary sample size to guarantee a prescribed level of confidence $100(1-\alpha)$ and margin of error B.

Sample size determination about μ

Example. In a biomedical experiment, we would like to estimate the population mean remaining life μ of healthy rats that are given a certain dose of a toxic substance. Suppose that we would like to write a 95 percent confidence interval for μ with a margin of error $B = 2$ days. From past studies, remaining rat lifetimes have been approximated by a normal distribution with standard deviation $\sigma = 8$ days. How many rats should we use for the experiment?

Sample size determination about μ

Example. In a biomedical experiment, we would like to estimate the population mean remaining life μ of healthy rats that are given a certain dose of a toxic substance. Suppose that we would like to write a 95 percent confidence interval for μ with a margin of error $B = 2$ days. From past studies, remaining rat lifetimes have been approximated by a normal distribution with standard deviation $\sigma = 8$ days. How many rats should we use for the experiment?

Solution Note that $z_{0.05/2} = 1.96, B = 2, \sigma = 8$, the desired sample size to estimate μ is

$$n = \left(\frac{z_{\alpha/2}\sigma}{B}\right)^2 = n = \left(\frac{1.96 \times 8}{2}\right)^2 \approx 61.46$$

We would sample $n = 62$ rats to achieve these goals.

Try to find the optimal n when $B = 3$.

Sample size determination about p

Setting: Suppose We would like to compute a $100(1-\alpha)$ percent confidence interval for a populaton proportion p, which is known as

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

What sample size n should we use? To determine the necessary sample size n, we need specify two pieces of information:

- the confidence level $100(1-\alpha)$
- the margin of error $B = z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$

However, the margin of error B depends on $\hat{p}$ that is obtained once we know the sample size n. To overcome this paradox, we replace $\hat{p}$ with p_0, a **priori guess** for p. If there is no sensible guess for p, use $p_0 = 0.5$. Therefore, the necessary sample size n is calculated by

$$B = z_{\alpha/2}\sqrt{\frac{p_0(1-p_0)}{n}} \Leftrightarrow n = \left(\frac{z_{\alpha/2}}{B}\right)^2 p_0(1-p_0)$$

Sample size determination about p

Example. You have been asked to estimate the proportion of raw material(in a certain manufacturing process) that is being "scrapped", that is, the material is so defective that it can not be reworked. If this proportion is larger than 10 percent, this will be deemed(by management) to be an unacceptable continued operating cost and a substantial process overhaul will be performed. Past experience suggests that the scrap rate is about 5 percent, but recent information suggests that this rate may be increasing. You would like to write a 95 percent confidence interval for p, the population proportion of raw material that is to be scrapped, with a margin of error equal to $B = 0.02$. How many pieces of material should you ask to ba sampled?

Sample size determination about p

Example. You have been asked to estimate the proportion of raw material(in a certain manufacturing process) that is being "scrapped", that is, the material is so defective that it can not be reworked. If this proportion is larger than 10 percent, this will be deemed(by management) to be an unacceptable continued operating cost and a substantial process overhaul will be performed. Past experience suggests that the scrap rate is about 5 percent, but recent information suggests that this rate may be increasing. You would like to write a 95 percent confidence interval for p, the population proportion of raw material that is to be scrapped, with a margin of error equal to $B = 0.02$. How many pieces of material should you ask to ba sampled?

Solution. Note that $z_{0.05/2} \approx 1.96$, we have

$$p_0 = 0.05 \Rightarrow n = \left(\frac{1.96}{0.02}\right)^2 (0.05)(1-0.05) \approx 457$$

$$p_0 = 0.1 \Rightarrow n = \left(\frac{1.96}{0.02}\right)^2 (0.1)(1-0.1) \approx 865$$