BUDT 730

Data, Models and Decisions

Lecture 6

Confidence Interval (II)

Prof. Sujin Kim

Agenda

- Produce a confidence interval for the mean with a certain level of precision
- Learn the properties of a point estimate of the population proportion
- Calculate and interpret a confidence interval for the proportion
- Produce a confidence interval for the population proportion with a certain level of precision
- Data Files:
  - FastFoodData.xlsx
  - Satisfaction Ratings.xlsx

CH8
Confidence Interval Estimation

Confidence Interval for **Mean** with **unknown** σ

Large sample size situation

- In the previous example, we use Z-statistics:
  - the **sample size is large** (≥ 30)
  - The **standard deviation is known**.

- What if the standard deviation is **unknown**?
  - Replace the standard deviation by the **sample standard deviation**
  - When this replacement is made, a new source of variability is introduced, and the sampling distribution is no longer normal.

- What if the sample size is small?
  - We must make an assumption on the underlying distribution to obtain a valid statistic – Normality assumption

T Statistic

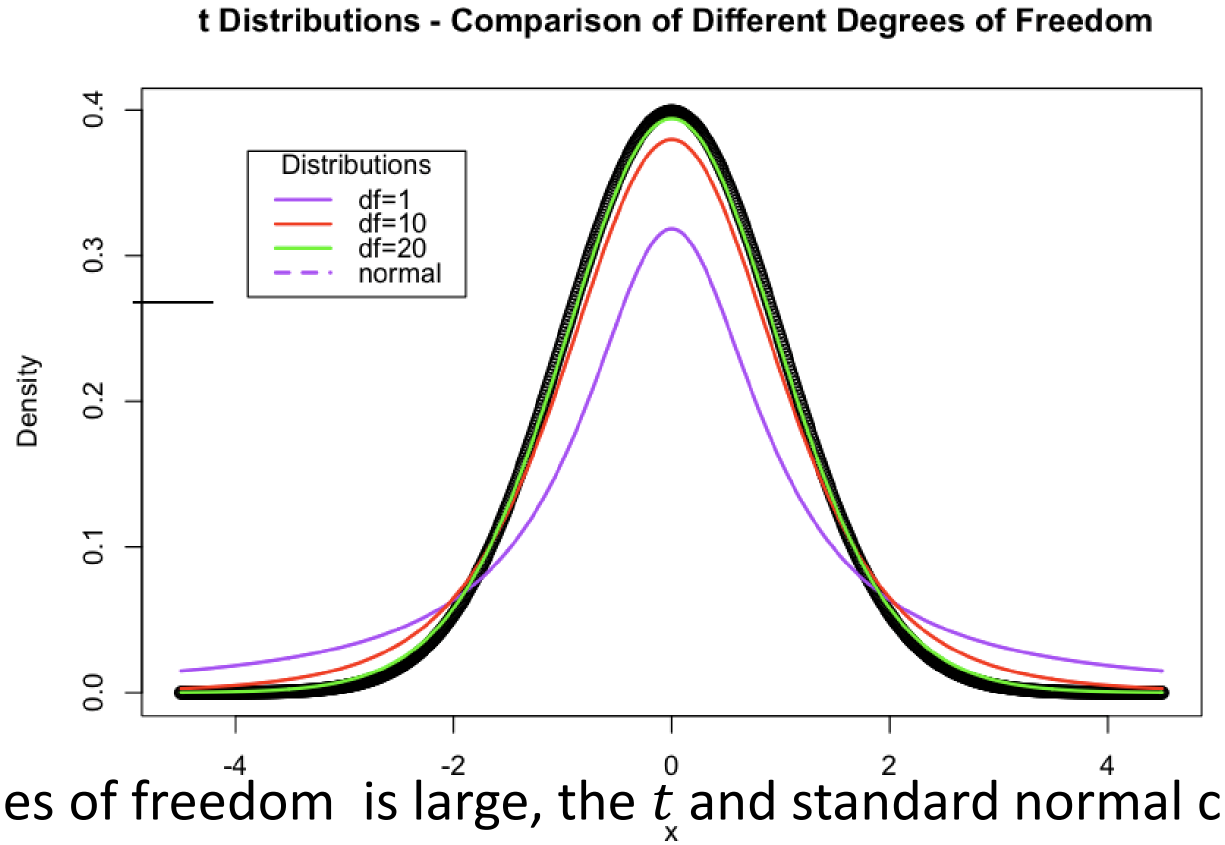
- σ is rarely known, so we estimate it with *s*, the sample standard deviation
- Assume that the population has a **normal** distribution with **unknown** standard deviation. Then, the T- statistic

$$\boxed{T = \frac{\bar{X} - \mu}{\boldsymbol{s}/\sqrt{n}}}$$

has a $\boldsymbol{t}$ **distribution with** $(\boldsymbol{n}-\mathbf{1})$ **degrees of freedom**.

- The appearance of the t distribution is similar to the standard normal distribution
- However, the t distribution has heavier tails than the normal; that is, it has more probability in the tails than the normal distribution
- As n gets larger, the t distribution gets closer to the standard normal distribution

Standard Normal vs. t Distribution



- When the degrees of freedom is large, the t and standard normal curves are practically the same.
- Thus, both the Z-test and T-test can be used to make inferences about a population mean with **unknown** standard deviation when the **sample size is large ($\geq$30)**.

CI and Hypothesis Tests for a Population Mean (9/29)

		Large Sample Size n	Small Sample Size
Known σ	Normal Population	**Z-statistic**	Z-statistic
	Non-Normal Population	**Z-statistic**	Cannot do
Unknown σ	Normal Population	**T-statistic** (or Z-statistic)	T-statistic
	Non-Normal Population	**T-statistic** (or Z-statistic)	Cannot do

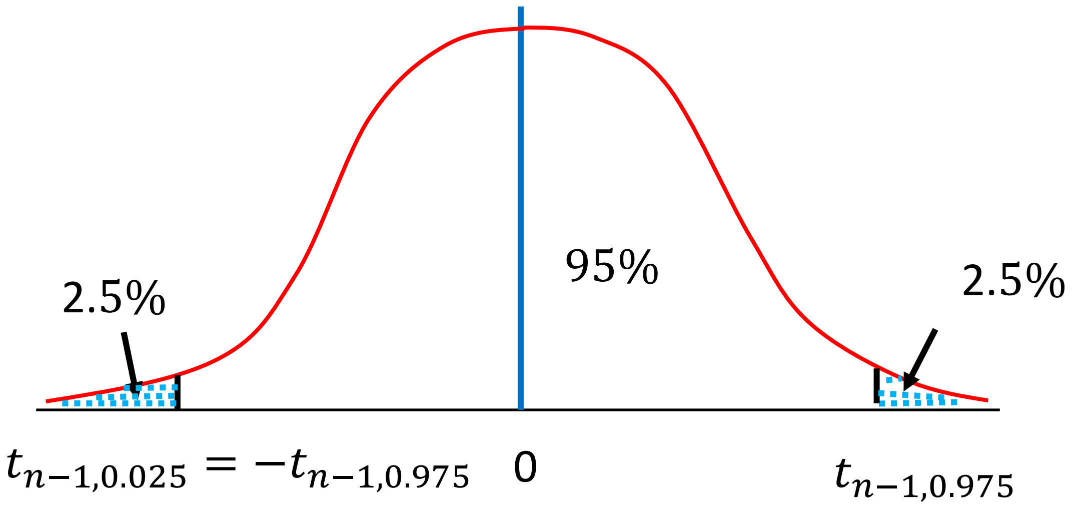
Confidence Interval for **Mean** with **unknown $\boldsymbol{\sigma}$**

- This is the equation for determining of the confidence interval for a sample mean when the standard deviation is unknown

$$\boxed{\bar{X} \pm (t - multiple) \times \frac{s}{\sqrt{n}}}$$

  - s: The sample standard deviation of the population
  - $t - multiple$: The t-value is determined by the confidence selected for the interval and the number of samples.

t-multiple



$$P\left(\bar{X} - t_{n-1,0.975}\frac{s}{\sqrt{n}} \le \mu \le \bar{X} + t_{n-1,0.975}\frac{s}{\sqrt{n}}\right) = 95\%$$

$$95\%\ \text{Confidence Interval} = \bar{X} \pm t_{n-1,0.975}\frac{s}{\sqrt{n}}$$

Generating t Values in Excel

- T.DIST function: Calculate probabilities
  - T.DIST(t, df, 1): P(T ≤ t), area to the left
  - T.DIST.RT(t, df): P(T ≥ t), area to the right
  - T.DIST.2T(t, df): P(T ≤ -t or T ≥t), area of two tails
- T.INV function: Calculate percentiles
  - T.INV(α, df): Find t such that P(T ≤ t) = α
    - $t_{n-1,\alpha}$= T.INV($\alpha, n-1$)
  - T.INV.2T(α, df): Find t such that P(T $\leq -t$, T $\geq t$) = α
    - **$t_{n-1,1-\frac{\alpha}{2}}$ = T.INV.2T($\alpha, n-1$)**

R functions for Z and t-multiples

- pnorm, qnorm, pt, and qt come standard with R.

```
>pnorm(1.96) (or pnorm(1.96, mean=0, sd=1))
[1] 0.9750021
> qnorm(0.975) (or pnorm(0.975, mean=0, sd=1))
[1] 1.959964
> qnorm(c(0.025,0.975), mean=0, sd=1)
[1] -1.959964  1.959964
> pt(2.0,29) (n=30 and df=n-1=29)
[1] 0.9725282
> qt(0.975,29)
[1] 2.04523
```

In-Class Exercise

FastFoodData.xlsx

Example: Fast Food Data

- The manager of a local fast-food restaurant is interested in improving the service provided to customers who use the restaurant's drive-up window.
- As a first step in this process, the manager asks his assistant to record **the time it takes to serve** a large number of customers at the final window in the facility's drive-up system.
- The time is measured in seconds.
- The results are in the file FastFoodData.xlsx.
- The file consists of 2 worksheets:
  - The first the worksheet "FullData" contains 1184 service times. For this problem you can assume that the population is the data in this worksheet.
  - The second worksheet is called "Sample". You can generate a random sample of 30 observations from the full data using the simple random sampling.

Example: Fast Food Data – Simple Random Sampling

1) Generate a new variable called "Random Number"
2) Populate the variable using function RAND() (uniformly distributed random number on (0,1))
3) Sort the data by this new variable
4) Select the first 30 observations
5) Copy the samples to a new sheet

A	B	C	D
Customer	Time	Random Number	
1	42	=RAND()	
2	35		
3	34		

Example: Fast Food Data

Using the information in `Sample' worksheet, calculate

- Sample mean =
- Population mean =
- Sampling error =
- Standard error =
- 95% Confidence interval =
- 90% Confidence interval =

Example: Fast Food Data - CI Calculation by Hand

$\text{n} = 30,$
$\text{sample mean} = 49.77,$
population mean = 55.45,
sample error = 49.77-55.45 = -5.89,
$\text{sample stdev} = 21.93$

Example: Fast Food Data - CI Calculation by Hand

Since $n \geq 30$, the normal sampling distribution can be justified.

100 (1 – α)% CI for the population mean μ

$$\bar{X} \pm t_{n-1,1-\frac{\alpha}{2}} \frac{s}{\sqrt{n}}$$

Standard error = $\frac{s}{\sqrt{n}} = \frac{21.93}{\sqrt{30}}$= 4.00

In Excel: $t_{29,\ 0.975}$ = T.INV(0.975, 29) or T.INV.2T(0.05,29) =2.05

In R: qt(0.975,29)=2.04523≈2.05

95% CI.: MOE = 2.05*4.00 = 8.2 ➔ 49.77 ± 8.2 ➔ (41.57, 57.97)

In Excel: $t_{29,\ 0.95}$ = = T.INV(0.95, 29) or T.INV.2T(0.1,29)= 1.70

In R: qt(0.95,29)=1.699≈ 1.70

90% CI.: MOE = 1.70*4.00 = 6.80 ➔ 49.77 ± 6.80 ➔ (42.97, 56.57)

Sample Size Selection for Estimation of the Mean

Controlling Confidence Interval Length

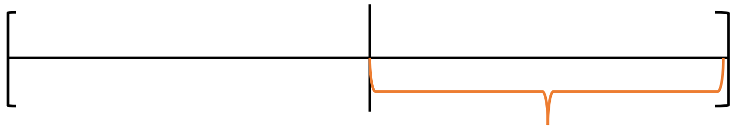
- The length of confidence interval is influenced by three things:
  - Variability in the population
  - Confidence level
  - Sample size
- What do we do if we want to produce a confidence interval with a certain level of precision (i.e., margin of error)?
  - We do not have any control over the variability in population.
  - The confidence level is typically set at 95%.
  - Therefore, the best way to control confidence interval is through the choice of **the sample size**.

Sample Size for Estimation of the Mean with Known σ

- Recall the formula for a confidence interval for the mean with known σ:

$$\bar{X} \pm (Z - multiple) \times \frac{\sigma}{\sqrt{n}}$$

Point estimator = $\bar{X}$



$$MOE = B$$

- The most obvious way to control confidence interval length is to choose the sample size (n) appropriately, which we can calculate from the formula for the confidence interval.
- The goal is to make the half-length of this interval (i.e. margin of error) equal to some prescribed value B.

Sample Size for Estimation of the Mean with Known σ

- Given some desired margin of error *B,*
  - The appropriate **sample size** for estimation of the mean is

$$B = z - multiple \times \frac{\sigma}{\sqrt{n}}$$

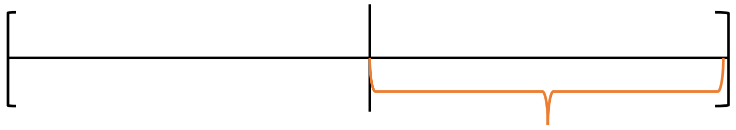
$$\Rightarrow \boxed{\boldsymbol{n = \left(\frac{z - multiple \times \sigma}{B}\right)^2}}$$

Sample Size for Estimation of the Mean with Unknown σ

- Recall the formula for a confidence interval for the mean with unknown σ:

$$\bar{X} \pm (t - multiple) \times \frac{s}{\sqrt{n}}$$

Point estimator = $\bar{X}$



$MOE = B$

Sample Size for Estimation of the Mean with Unknown σ

- Given some desired margin of error *B,*
  - The appropriate sample size for estimation of the mean is

$$B = t - multiple \times \frac{s}{\sqrt{n}}$$

$$\Rightarrow \quad n = \left(\frac{t - multiple \times s}{B}\right)^2$$

Here, s is the sample standard deviation.

Sample Size for Estimation of the Mean with Unknown σ

- Unfortunately, sample size selection must be done before a sample is observed, so "$\boldsymbol{s}$" is not yet available
  - Replace $\boldsymbol{s}$ by some reasonable estimate $\boldsymbol{\sigma_{est}}$ of the population standard deviation $\boldsymbol{\sigma}$.
  - *n* also affects *t-multiple*, so we can use z-multiple instead
    - When the sample size is large, z-values and t-values are practically equal
- The resulting **sample size formula** is:

$$\boxed{\boldsymbol{n = \left(\frac{Z - multiple * \sigma_{est}}{B}\right)^2}}$$

  - Typically, we round *n* up to the next larger integer.

Example: Meal Service Sample Size

- We wish to construct a 95% confidence interval for the mean number of meals served with a margin of error of 500 meals. How many samples should be collected to accomplish this?
  - The standard deviation is 1643.17
  - For 95% confidence level, the Z-multiple is 1.96

Example: Meal Service Sample Size

- We wish to construct a 95% confidence interval for the mean number of meals served with a margin of error of 500 meals. How many samples should be collected to accomplish this?
    - σ =1643.17
    - Z-multiple =1.96
    - B= 500
    - Plugging into the formula, we have

$$n = \left(\frac{Z - multiple * \sigma}{B}\right)^2 = \left(\frac{1.96 * 1643.17}{500}\right)^2 = 41.49 \approx 42$$

Example: Fast-Food Data

- Recall the summary statistics and a 95% confidence interval for the sample mean:
  - Sample mean 49.77
  - Sample Standard Deviation 21.93
  - Sample Size 30
  - MOE: 8.19
  - 95% CI 41.58 57.96

- If we want to reduce the size of the confidence interval by half, without changing the confidence (95%), how many samples would we need?

Example: Fast-Food Data

- σ_{est} = $s_{30} = 21.93$
- Z-multiple = 1.96
- B= 8.19 /2 ≈4.1
- $n \geq \left(\frac{1.96*21.93}{4.1}\right)^2 = 110$

CH8
Confidence Interval Estimation

For a Population Proportion

Example: Satisfaction Ratings

- File: Satisfaction Ratings.xlsx
- A fast-food manager restaurant added a new sandwich to its menu.
- A random sample of 40 customers who ordered a new sandwich were surveyed.
- Each of these customers was asked to rate the sandwich on a scale of 1 to 10, 10 being the best
- The manager would like to estimate the proportion (p) of customers who rate the new sandwich at least 6
    - This estimates the proportion of people who like the new sandwich

Customer	Satisfaction
1	7
2	5
3	5
4	6

Point Estimate for a **Proportion**

- Very similar to the procedure for a population mean
- Let p be the proportion of the population with property A
- From a random sample of size n, the sample proportion is

$$\boxed{\begin{aligned}\hat{p} &= \text{\# of observations of interest /sample size (n)}\\ &= \frac{\sum_{i=1}^{n} p_i}{n}\end{aligned}}$$

$$p_i = \begin{cases} 1 & if\ the\ ith\ member\ has\ property\ A \\ 0 & otherwise \end{cases}$$

- Then, $\hat{p}$ is the point estimate of the population proportion p.

Sampling distribution of the sample proportion

- Note that $\sum_{i=1}^{n} p_i \sim Binomial\ (n, p)$
- Normal approximation to Binomial:
  - For large n, $Binomial\ (n, p)$ is approximately normally distributed.
- For sufficiently large n ($\boldsymbol{np}$ **&** $\boldsymbol{n(1-p) \geq 5}$), the sampling distribution of $\hat{p}$ is approximately normal with mean p and standard error $\sqrt{\frac{p(1-p)}{n}}$:

$$\boxed{\hat{p} \sim Normal\left(p, \sqrt{\frac{p(1-p)}{n}}\right)}$$

Confidence Interval for a Proportion

- p is unknown. Thus, $\hat{p}$ is substituted for p in this standard error.
- The estimated standard error of sample proportion is

$$SE(\hat{p}) \approx \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

- The multiple used to obtain a confidence interval is a Z-multiple
- CI for a proportion:

$$\boxed{\hat{p} \pm (Z-multiple) \times \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}}$$

Example: Satisfaction Ratings - Solution by Hand

- Sample size: n = 40
- Data processing: Add a column of dummy variable indicating whether each person rated the sandwich at least 6
  - IF('cell'>=6, 1, 0))
- **25 customers rate the sandwich at least 6**
- Sample proportion:
- Estimated standard error:
- Desired confidence = 95% → Z-multiple =
- 95% CI =

Customer	Satisfaction	At least 6?
1	7	1
2	5	0
⋮	⋮	⋮

Example: Satisfaction Ratings - Solution by Hand

- Sample size: n = 40
- Data processing: Add a column of dummy variable indicating whether each person rated the sandwich at least 6
  - IF('cell'>=6, 1, 0))
- **25 customers rate the sandwich at least 6**
- Sample proportion: $\hat{p}$ =25/40 = 0.625
- Estimated standard error: $\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$ = 0.07655
- Desired confidence = 95% → Z-multiple = $z_{0.975}$ = 1.96

$$\hat{p} \pm (Z - multiple) \times \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

$\Rightarrow$ 0.625±1.96*(0.07655) =0.625±(0.150)=(0.475, 0.775)

Example: Satisfaction Ratings

- 95% CI is (47.5%, 77.5%)
- Interpretation:

  "Based on this sample of size 40, the manager can be 95% confident that the percentage of all customers who would rate the sandwich 6 or higher is between 47.5% and 77.5%"
- This CI is very wide, so there is still a lot of uncertainty about the true population proportion.
- To reduce the length of this interval, the manager would need to sample more customers.

Sample Size Selection
for Estimation of the Proportion

Example: Satisfaction Ratings

- 25 customers rate the sandwich at least 6 : $\hat{p} = 0.625$ =62.5%
- The 95% CI is (47.5%, 77.5%)
- This CI is very wide, so there is still a lot of uncertainty about the true population proportion.
- The manager would like to obtain an estimate of the proportion that is accurate to within 3% or 0.03 with 95% confidence.
- **Then, how large a sample size should the company use to achieve this?**

Sample Size for Estimation of the Proportion

- Recall the CI formula for a proportion and the margin of error (MOE) is

$$MOE = (Z - multiple) \times \sqrt{\frac{p(1-p)}{n}}$$

- Given some desired margin of error *B*,

$$B = (Z - multiple) \times \sqrt{\frac{p(1-p)}{n}}$$

- The population proportion p is unknown, and we replace p by some reasonable estimate p_{est} .

$$\boxed{n = \left(\frac{Z - multiple}{B}\right)^2 p_{est}(1 - p_{est})}$$

Example: Satisfaction Ratings

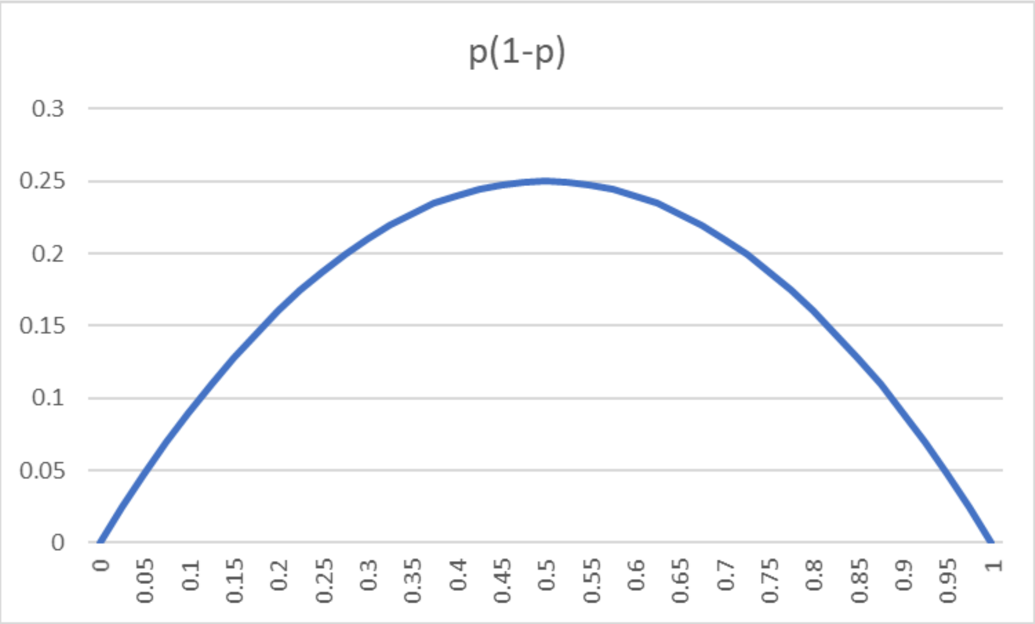
Q: How large should the sample be if they would like to *ensure* that MOE (=*B*) is no larger than 0.03 or 3% at 95% confidence?

- Prior to survey, the manager has no knowledge about p.

$$n = \left(\frac{Z\text{-}multiple}{B}\right)^2 p_{est}(1 - p_{est})$$

How to select p_{est}?

Note that $p(1-p)$ is maximized when p is 0.5.



Option 1: No prior information on $\boldsymbol{p}$.

One approach is to **use the worst-case scenario and use $\boldsymbol{p_{est} = 0.5}$**

Example: Satisfaction Ratings

Q: How large should the sample be if they would like to *ensure* that MOE (=*B*) is no larger than 0.03 or 3% at 95% confidence?

We have

p_{est} = 0.5

Z-multiple = 1.96

B= 0.03

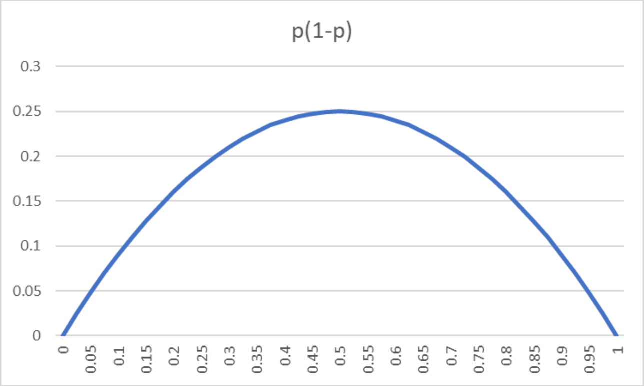
$$n = \left(\frac{Z\!-\!multiple}{B}\right)^2 p_{est}(1 - p_{est}) = \left(\frac{1.96}{0.03}\right)^2 (0.5)(0.5) \approx 1067$$

Conclusion: To obtain a 95% CI of this length (**3%**) for a population proportion, **only about 1000 people** need to be sampled, **regardless of the population size**.

How to Select $\boldsymbol{p_{est}}$?

Option 2: Prior information on p $- range\ of\ p$

- If the manager has a prior knowledge about p, say p is between p_L and p_U. We can select a value that gives the most conservative n.
  - For example, if p will likely be somewhere between 0.1 and 0.2 (that is between 10% and 20%), we can use $p_{est} = 0.2$.



Option 3: Prior information on p $-$ p_{est} is given

- We can also collect a small number of samples to estimate it:
  - Ex: Based on 40 customers, $p_{est} = 0.625$

$$n = \left(\frac{Z\text{−}multiple}{B}\right)^2 p_{est}(1 - p_{est}) = \left(\frac{1.96}{0.03}\right)^2 (0.625)(0.375) = 1000.4 \approx 1001$$

Example: Fast-Food Data (10/4(M))

- Using the information in the sample worksheet,
  - Estimate **the probability that the service time is greater than one minute (By hand and Excel, By R)**
  - Compute the 90% confidence interval of the probability.
  - Suppose that the probability will likely be somewhere between 0.1 and 0.4. How large should the sample be if they would like to *ensure* that MOE (=*B*) is no larger than 0.05 or 5% at 90% confidence?

Example: Fast-Food Data

- $n = 30, \hat{p} = \frac{6}{30} = 0.2$
- $\hat{p} \pm (Z - multiple) \times \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = 0.2 \pm (1.645) * \sqrt{\left(\frac{0.2*0.8}{30}\right)} = 0.2 \pm 0.120134$
- $n = \left(\frac{Z-multiple}{B}\right)^2 p_{est}(1 - p_{est}) = \left(\frac{1.645}{0.05}\right)^2 (0.4)(0.6) \approx 260$

Section 8.7: Confidence Interval for the Difference Between Means (You may skip this topic)

The comparison of two population means

- We would like to compare the difference between

$$\mu_A = \mathrm{E}(X_A) \text{ and } \mu_B = \mathrm{E}(X_B)$$

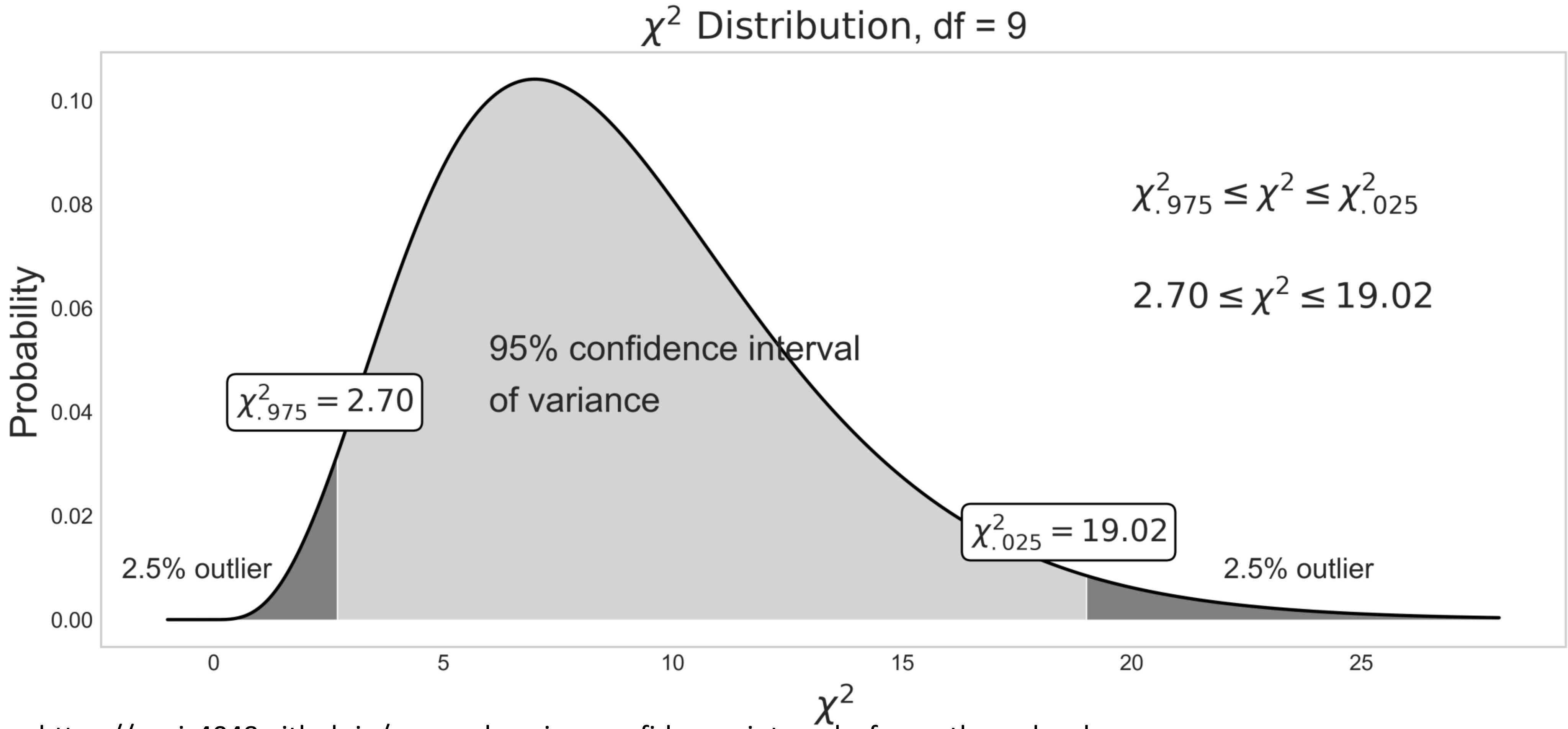
  - For statistical reasons, we need to distinguish when comparing **independent samples** to **paired samples**
  - The procedure for paired samples is very straightforward: Analyze $\mathrm{d} = X_A - X_B$
  - For independent samples, we need to handle two independent samples.
- We will discuss the details of two sample analysis in Ch 9.

Section 8.6: Confidence Interval for a Standard Deviation (You may skip this topic)

- There are cases where the variability in the population, measured by σ, is of interest in its own right.
  - The sample standard deviation s is used as a point estimate of σ.
  - However, the sampling distribution of s is not symmetric—it is not the normal distribution or the *t* distribution.
  - The appropriate sampling distribution is a right-skewed distribution called the **chi-square distribution**.
    - Like the *t* distribution, the chi-square distribution has a degrees of freedom parameter.
- 95% confidence interval on $\sigma^2\ is$

$$\frac{(n-1)s^2}{\chi^2_{0.025,n-1}} \le \sigma^2 \le \frac{(n-1)s^2}{\chi^2_{0.975,n-1}}$$

χ^2 Distribution



Source: https://aegis4048.github.io/comprehensive_confidence_intervals_for_python_developers

Next …

- Ch9 Hypothesis testing