# Chapter 1. Statistical inference in one population

## Contents



# Chapter 1. Statistical inference in one population

## Learning goals

At the end of this chapter you should know how to:

- Estimate the unknown population parameters from the sample data
- Construct confidence intervals for the unknown population parameters from the sample data:
  - In the case of a normal distribution: confidence intervals for the population mean and variance
  - In large samples: confidence intervals for the population mean and proportion
- Interpret the confidence interval
- Understand the impact of the sample size, confidence level, etc on the length of the confidence interval
- Calculate a sample size needed to control a given interval width

## Chapter 1. Statistical inference in one population

### References

- Newbold, P. "Statistics for Business and Economics"
  - Chapters 7 and 8 (8.1-8.6)
- Ross, S. "Introduction to Statistics"
  - Chapter 8

## Statistical inference: key words (i)

- Population: the complete set of numerical information on a particular quantity in which an investigator is interested.
  - We identify the concept of the population with that of the random variable $X$.
  - The law or the distribution of the population is the distribution of $X$, $F_X$.
- Sample: an observed subset (say, of size $n$) of the population values.
  - Represented by a collection of $n$ random variables $X_1, X_2, \ldots, X_n$, typically iid (independent identically distributed).
- Parameter: a constant characterizing $X$ or $F_X$.

# Statistical inference: key words (ii)

- Statistical inference: the process of drawing conclusions about a population on the basis of measurements or observations made on a sample of individuals from the population.
- Statistic: a random variable obtained as a function of a random sample, $X_1, X_2, \dots, X_n$
- Estimator of a parameter: a random variable obtained as a function, say $T$, of a random sample, $X_1, X_2, \dots, X_n$, used to estimate the unknown population parameter.
- Estimate: a specific realization of that random variable, i.e., $T$ evaluated at the observed sample, $x_1, x_2, \dots, x_n$, that provides an approximation to that unknown parameter.

# Statistical inference: example

| | | | | |
|---|---|---|---|---|
| We want to know $\mu_X = \mathrm{E}[X]$ | | We have $n$ copies of $X$ | | We have $n$ observed values of $X_1, X_2, \dots, X_n$ |
| $X \sim F$ | $\Rightarrow$ | $X_1, X_2, \dots, X_n \sim F$<br>Sample | $\Rightarrow$ | $x_1, x_2, \dots, x_n$<br>Observed sample |
| $\Downarrow$ | | $\Downarrow$ | | $\Downarrow$ |
| $\mu_X = \mathrm{E}[X]$<br>Expected value of $X$ | $\Leftarrow$ | Estimator of $\mu_X$ (r. v.)<br>$\bar{X}$<br>Sample mean | $\Leftarrow$ | Estimate of $\mu_X$ (number)<br>$\bar{x}$<br>Sample mean |

# Point estimators: introduction

- A point estimator of a population parameter is a function, call it $T$, of the sample information $\underline{X}_n = (X_1, \ldots, X_n)$ that yields a single number.
- Examples of population parameters, estimators and estimates:

| Population parameter | $T(\underline{X}_n)$ | Estimator: notation | Estimate: notation |
|---|---|---|---|
| Pop. mean $\mu_X$ | sample mean $\frac{X_1+\ldots+X_n}{n}$ | $\bar{X} = \hat{\mu}_X$ | $\bar{x}$ |
| Pop. prop. $p_X$ | sample prop. | $\hat{p}_X$ | $\hat{p}_x$ |
| Pop. var. $\sigma^2_X$ | sample var. $\frac{\sum_i X_i^2 - n(\bar{X})^2}{n}$ | $\hat{\sigma}^2_X$ | $\hat{\sigma}^2_x$ |
| Pop. var. $\sigma^2_X$ | sample quasi. var. $\frac{\sum_i X_i^2 - n(\bar{X})^2}{n-1} = \frac{n}{n-1}\hat{\sigma}^2_X$ | $s^2_X$ | $s^2_x$ |
| ... | ... | ... | ... |
| In general, $\theta_X$ | ... | $\hat{\theta}_X$ | $\hat{\theta}_x$ |

# Point estimators: properties (i)

What are desirable characteristics of the estimators?

- Unbiasdness. This means that the bias of the estimator is zero. What's bias? Bias equals the expected value of the estimator minus the target parameter

$$\text{Bias}[\hat{\theta}_X] = \text{E}[\hat{\theta}_X] - \theta_X$$

| Population parameter | Estimator $T(\underline{X}_n)$ | Bias | Unbiased? | Minimum Variance Unbiased Estimator? |
|---|---|---|---|---|
| Pop. mean $\mu_X$ | $\bar{X}$ | $\text{E}[\bar{X}] - \mu_X = 0$ | Yes | Yes, if $X$ normal |
| Pop. prop. $p_X$ | $\hat{p}_X$ | $\text{E}[\hat{p}_X] - p_X = 0$ | Yes | Yes |
| Pop. var. $\sigma^2_X$ | $\hat{\sigma}^2_X$ | $\text{E}[\hat{\sigma}^2_X] - \sigma^2_X \neq 0$ | No | No |
| Pop. var. $\sigma^2_X$ | $s^2_X$ | $\text{E}[s^2_X] - \sigma^2_X = 0$ | Yes | Yes, if $X$ normal |
| In general, $\theta_X$ | $\hat{\theta}_X$ | $\text{E}[\hat{\theta}_X] - \theta_X$ | Often | Rarely |

# Point estimators: properties (ii)

- Efficiency. Measured by the estimator's variance. Estimators with smaller variance are more efficient.
- Relative efficiency of two unbiased estimators $\hat{\theta}_{X,1}$ and $\hat{\theta}_{X,2}$ of a parameter $\theta_X$ is

$$\text{Relative efficiency}(\hat{\theta}_{X,1}, \hat{\theta}_{X,2}) = \frac{\text{Var}[\hat{\theta}_{X,1}]}{\text{Var}[\hat{\theta}_{X,2}]}$$

  Note:
  - sometimes the inverse is used as a definition
  - in any case, an estimator with smaller variance is more efficient

# Point estimators: properties (iii)

- A more general criterion to select estimators (among unbiased and biased ones) is the mean squared error defined as

$$\text{MSE}[\hat{\theta}_X] = \text{E}[(\hat{\theta}_X - \theta_X)^2] = \text{Var}[\hat{\theta}_X] + (\text{Bias}[\hat{\theta}_X])^2$$

  Note:
  - the mean squared error of an unbiased estimator equals its variance
  - an estimator with smaller MSE is better
  - the minimum variance unbiased estimator has the smallest variance/MSE among all estimators
- How do we come up with the definition of the estimator $T$?
  - In some situations, there exists an optimal estimator called minimum variance unbiased estimator.
  - If that's not the case, there are various alternative methods that yield reasonable estimators, for example:
    - Maximum likelihood estimation
    - Method of moments

# Point estimation: example

**Example:** 7.1 (Newbold) Price-earnings ratios for a random sample of ten stocks traded on the NY Stock Exchange on a particular day were

$$10,\ 16,\ 5,\ 10,\ 12,\ 8,\ 4,\ 6,\ 5,\ 4$$

Use an unbiased estimation procedure to find point estimates of the following population parameters: mean, variance, proportion of values exceeding 8.5.

$$\begin{aligned} \bar{x} &= \frac{80}{10} = 8 \\ s_x^2 &= \frac{782 - 10(8)^2}{10-1} = 15.78 \\ \hat{p}_x &= \frac{1+1+0+1+1+0+0+0+0+0}{10} \\ &= 0.4 \end{aligned}$$

# Point estimation: example

**Example:** Let $\hat{\mu}_X = \frac{2}{n(n+1)}(X_1 + 2X_2 + \ldots + nX_n)$ be an estimator of the population mean based on a SRS $\underline{X}_n$. Compare this estimator with the sample mean, $\bar{X}$.

We know that $\bar{X}$ is an unbiased estimator of $\mu_X$, whose variance is $\frac{\sigma_X^2}{n}$.

$\hat{\mu}_X$ is also unbiased:

$$\begin{aligned} E[\hat{\mu}_X] &= E\left[\frac{2}{n(n+1)}(X_1 + 2X_2 + \ldots + nX_n)\right] \\ &= \frac{2}{n(n+1)}(E[X_1] + 2E[X_2] + \ldots + nE[X_n]) \\ &=_{id} \frac{2}{n(n+1)}(\mu_X + 2\mu_X + \ldots + n\mu_X) \\ &= \frac{2\mu_X}{n(n+1)}\overbrace{(1+2+\ldots+n)}^{n(n+1)/2} = \mu_X \end{aligned}$$

$\Rightarrow$ *Bias*$[\hat{\mu}_X] = 0$

And its variance/MSE is:

$$\begin{aligned} V[\hat{\mu}_X] &= V\left[\frac{2}{n(n+1)}(X_1 + 2X_2 + \ldots + nX_n)\right] \\ &=_{indep.} \left(\frac{2}{n(n+1)}\right)^2 (V[X_1] + 2^2V[X_2] + \ldots + n^2V[X_n]) \\ &=_{id} \frac{4}{n^2(n+1)^2}\sigma_X^2 \overbrace{(1^2 + 2^2 + \ldots + n^2)}^{n(n+1)(2n+1)/6} \\ &= \frac{2(2n+1)}{3n(n+1)}\sigma_X^2 \end{aligned}$$

$$MSE[\hat{\mu}_X] = V[\hat{\mu}_X] + 0^2 = \frac{2(2n+1)}{3n(n+1)}\sigma_X^2$$

$$\text{Relative efficiency}(\bar{X}, \hat{\mu}_X) = \frac{\sigma_X^2/n}{\frac{2(2n+1)}{3n(n+1)}\sigma_X^2} = \frac{3(n+1)}{2(2n+1)}$$

It's easy to see that for $n \geq 2$, this ratio is smaller than 1 so $\bar{X}$ is a more efficient estimator for $\mu_X$.

## From point estimation to confidence interval estimation

- So far, we have consider the point estimation of an unknown population parameter which, assuming we had a SRS sample of $n$ observations from $X$, would produce an educated guess about that unknown parameter
- Point estimates however, do not take into account the variability of the estimation procedure due to, among other factors:
  - sample size - surely, larger samples should provide more accurate information about the population parameter
  - variability in the population - samples from populations with smaller variance should give more accurate estimates
  - whether other population parameters are known
  - etc

These drawbacks can be overcome by considering confidence interval estimation, that is, a method that gives a range of values (an interval) in which the parameter is likely to fall.

## Confidence interval estimator and confidence interval

Let $\underline{X}_n = (X_1, X_2, \ldots, X_n)$ be a SRS from a population $X$ with a cdf $F_X$ that depends on an unknown parameter $\theta$.

- A confidence interval estimator for $\theta$ at a confidence level $(1-\alpha) = 100(1-\alpha)\%$ is an interval $(T_1(\underline{X}_n), T_2(\underline{X}_n))$ that satisfies

$$P\left(\theta \in (T_1(\underline{X}_n), T_2(\underline{X}_n)\right) = 1 - \alpha$$

  - Interpretation: we have a probability of $(1-\alpha)$ that the unknown population parameter will be in $(T_1(\underline{X}_n), T_2(\underline{X}_n))$.
- A confidence interval for $\theta$ at a confidence level $1-\alpha$ is the observed value of the confidence interval estimator,

$$(T_1(\underline{x}_n), T_2(\underline{x}_n))$$

  - Interpretation: we can be $(1-\alpha)$ confident that the unknown population parameter will be in $(T_1(\underline{x}_n), T_2(\underline{x}_n))$.

Typical levels of confidence

| $\alpha$ | 0.01 | 0.05 | 0.10 |
|---|---|---|---|
| $100(1-\alpha)\%$ | 99% | 95% | 90% |

# Finding confidence interval estimators: procedure

1. Find a quantity involving the unknown parameter $\theta$ and the sample $\underline{X}_n$, $C(\underline{X}_n, \theta)$, whose distribution is known and does not depend on the parameter - a so-called pivotal quantity or a pivot for $\theta$
2. Use the upper $1 - \alpha/2$ and $\alpha/2$ quantiles of that distribution and the definition of the confidence interval estimator to set up the equation

$$P(\overbrace{1 - \alpha/2 \text{ quantile} < C(\underline{X}_n, \theta) < \alpha/2 \text{ quantile}}^{\text{double inequality}}) = 1 - \alpha$$

3. To find the end points $T_1(\underline{X}_n)$ and $T_2(\underline{X}_n)$ of the confidence interval estimator, solve the double inequality for the parameter $\theta$
4. A $100(1 - \alpha)\%$ confidence interval for $\theta$ is $(T_1(\underline{x}_n), T_2(\underline{x}_n))$

# Confidence interval for the population mean, normal population with known variance

1. Let $\underline{X}_n$ be a SRS of size $n$ from $X$. Under the assumptions:
   - $X$ follows a normal distribution with parameters $\mu_X$ and $\sigma_X^2$
   - $\sigma_X^2$ is known (rather unrealistic)
2. The pivotal quantity for $\mu_X$ is

$$\boxed{\frac{\bar{X} - \mu_X}{\sigma_X/\sqrt{n}} \sim N(0, 1)}$$

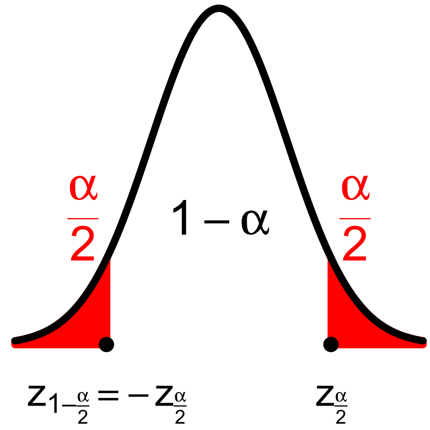
   - Note: the standard deviation of $\bar{X}$, $\sigma_X/\sqrt{n}$, (or any other stats) is called the standard error

# Confidence interval for the population mean, normal population with known variance

3. Hence, if $z_{1-\alpha/2}$ and $z_{\alpha/2}$ are the $(1-\alpha/2)$ and $(\alpha/2)$ upper quantiles of the $N(0,1)$, we have

$$\boxed{P(z_{1-\alpha/2} < Z < z_{\alpha/2}) = 1-\alpha}$$

Standard normal density




Recall: If $Z \sim N(0,1)$ then $\mathrm{E}[Z]=0$, $\mathrm{V}[Z]=1$

4. Therefore $P(\overbrace{z_{1-\alpha/2}}^{-z_{\alpha/2}} < \overbrace{\frac{\bar{X}-\mu_X}{\sigma_X/\sqrt{n}}}^{Z} < z_{\alpha/2}) = 1-\alpha$

# Confidence interval for the population mean, normal population with known variance

5. Solve the double inequality for $\mu_X$:

$$\begin{array}{rcl} -z_{\alpha/2} & < \frac{\bar{X}-\mu_X}{\sigma_X/\sqrt{n}} < & z_{\alpha/2} \\ -z_{\alpha/2}\frac{\sigma_X}{\sqrt{n}} & < \bar{X}-\mu_X < & z_{\alpha/2}\frac{\sigma_X}{\sqrt{n}} \\ -z_{\alpha/2}\frac{\sigma_X}{\sqrt{n}}-\bar{X} & < -\mu_X < & -\bar{X}+z_{\alpha/2}\frac{\sigma_X}{\sqrt{n}} \\ z_{\alpha/2}\frac{\sigma_X}{\sqrt{n}}+\bar{X} & > \mu_X > & \bar{X}-z_{\alpha/2}\frac{\sigma_X}{\sqrt{n}} \end{array}$$

to obtain the confidence interval estimator

$$(\overbrace{\bar{X}-z_{\alpha/2}\frac{\sigma_X}{\sqrt{n}}}^{T_1(\underline{X}_n)}, \overbrace{\bar{X}+z_{\alpha/2}\frac{\sigma_X}{\sqrt{n}}}^{T_2(\underline{X}_n)})$$

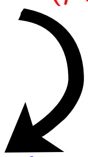
6. The confidence interval is:

$$\mathrm{CI}_{1-\alpha}(\mu_X) = \left(\bar{x}-z_{\alpha/2}\frac{\sigma_X}{\sqrt{n}}, \bar{x}+z_{\alpha/2}\frac{\sigma_X}{\sqrt{n}}\right) = \left(\bar{x}\mp z_{\alpha/2}\frac{\sigma_X}{\sqrt{n}}\right)$$

# Example: finding a confidence interval for $\mu_X$

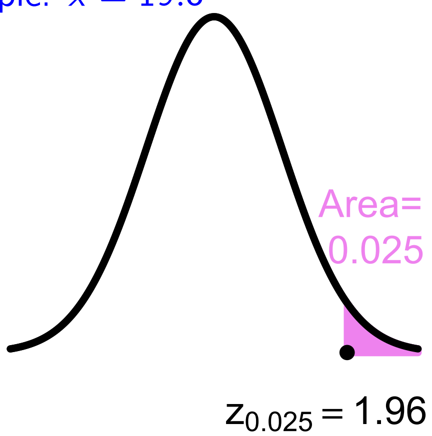
**Example:** 8.2 (Newbold) A process produces bags of refined sugar. The weights of the contents of these bags are normally distributed with standard deviation 1.2 ounces. The contents of a random sample of twenty-five bags had mean weight 19.8 ounces. Find a 95% confidence interval for the true mean weight for all bags of sugar produced by the process.

Population:
$X$ = "weight of a sugar bag (in oz)"
$X \sim N(\mu_X, \sigma_X^2 = 1.2^2)$



SRS: $n = 25$

Sample: $\bar{x} = 19.8$




Objective: $CI_{0.95}(\mu_X) = \left(\bar{x} \mp z_{\alpha/2}\frac{\sigma_X}{\sqrt{n}}\right)$

$$\sigma_X = 1.2$$
$$n = 25 \qquad \bar{x} = 19.8$$
$$1-\alpha = 0.95 \Rightarrow \alpha/2 = 0.025$$
$$z_{\alpha/2} = z_{0.025} = 1.96$$
$$CI_{0.95}(\mu_X) = \left(19.8 \mp 1.96\frac{1.2}{\sqrt{25}}\right) = (19.8 \mp 0.47) = (19.33, 20.27)$$

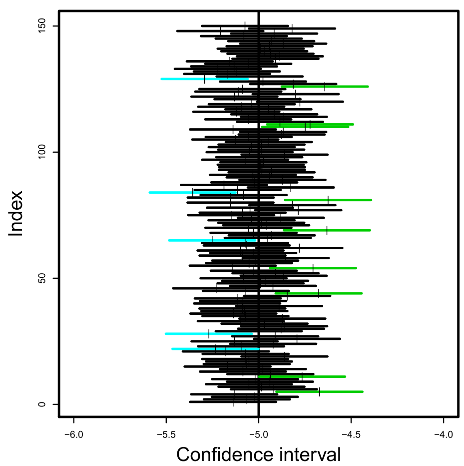
Interpretation: We can be 95% confident that $\mu_X$ is in $(19.33, 20.27)$

# Frequency interpretation of the CI, conf. level effect

In this simulated example, 150 samples of the same size $n = 50$ were generated from $X \sim N(\mu_X = -5, \sigma_X^2 = 1^2)$ and 150 $CI_{1-\alpha}(\mu_X)$ were constructed with $\alpha = 0.1$ and $\alpha = 0.01$.

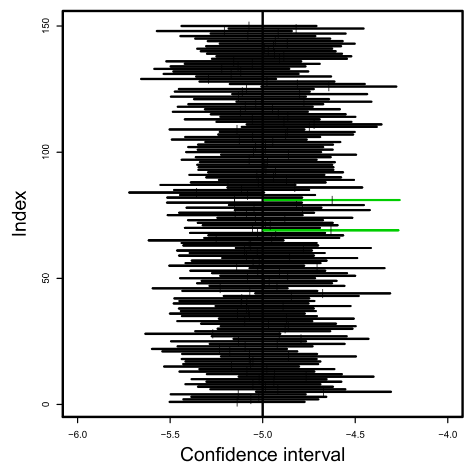
$\mu_X$ in approximately $150(0.9) = 135$ ints.
(but not in $150(0.1) = 15$)

$(1-\alpha) = 0.9$, $n = 50$




$\mu_X$ in approximately $150(0.99) = 148.5$ ints.
(but not in $150(0.01) = 1.5$)

$(1-\alpha) = 0.99$, $n = 50$



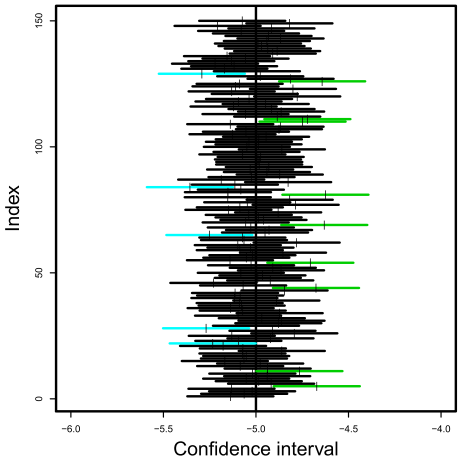

The width of the interval, $w = \bar{x} + \frac{z_{\alpha/2}\sigma_X}{\sqrt{n}} - \left(\bar{x} - \frac{z_{\alpha/2}\sigma_X}{\sqrt{n}}\right) = 2\frac{z_{\alpha/2}\sigma_X}{\sqrt{n}}$, increases with the increasing confidence level (keeping everything else the same). Why?

# Frequency interpretation of the CI, sample size effect

Here we collect 150 samples of size $n = 50$ and another 150 of size $n = 200$ from $\boxed{X \sim N(\mu_X = -5, \sigma_X^2 = 1^2)}$.

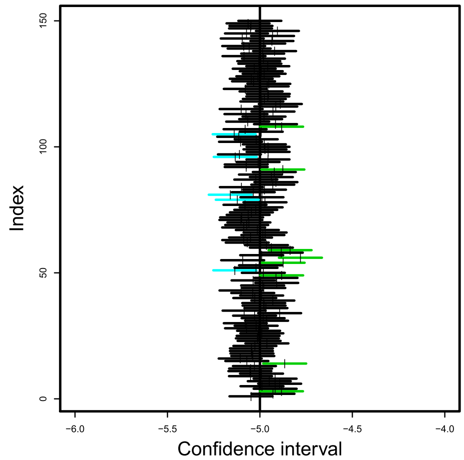
$\mu_X$ in approximately $150(0.9) = 135$ ints.
(but not in $150(0.1) = 15$)

$\boxed{(1-\alpha) = 0.9, n = 50}$




$\mu_X$ in approximately $150(0.9) = 135$ ints.
(but not in $150(0.1) = 15$)

$\boxed{(1-\alpha) = 0.9, n = 200}$




The width of the interval decreases with the increasing sample size (keeping everything else the same). Why?

Question: What is the effect of $\sigma$ on the width?

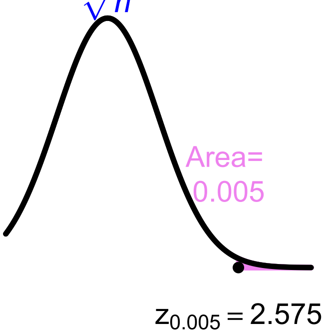
# Example: estimating the sample size

**Example:** 8.14 (Newbold) The lengths of metal rods produced by an industrial process are normally distributed with standard deviation 1.8mm. Suppose that a production manager requires a 99% confidence interval extending no further than 0.5mm on each side of the sample mean. How large a sample is needed to achieve such an interval?

Population:
$X =$ "length of a metal rod (in mm)"
$X \sim N(\mu_X, \sigma_X^2 = 1.8^2)$

SRS: $n = ?$

$CI_{0.99}(\mu_X)$: $\overbrace{2\frac{z_{\alpha/2}\sigma_X}{\sqrt{n}}}^{\text{width}} \leq 2(0.5) = 1$




Objective: $n$ such that width $\leq 1$

$$2\frac{z_{\alpha/2}\sigma_X}{\sqrt{n}} \leq 1$$
$$2z_{\alpha/2}\sigma_X \leq \sqrt{n}$$
$$85.93 = (2(2.575)(1.8))^2 \leq n$$

To satisfy the manager's requirement, a sample of at least 86 observations is needed.

## Confidence interval for the population mean in large samples

1. Let $\underline{X}_n$ be a SRS of size $n$ from $X$. Under the assumptions:
   - $X$ follows a nonnormal distribution with parameters $\mu_X$ and $\sigma_X^2$
   - the sample size $n$ is large ($n \geq 30$)
2. The pivotal quantity for $\mu_X$ based on the Central Limit Theorem is

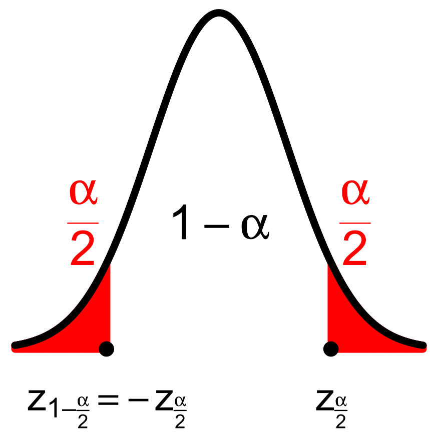
$$\boxed{\frac{\bar{X} - \mu_X}{\hat{\sigma}_X/\sqrt{n}} \sim_{\text{approx.}} N(0,1)}$$

## Confidence interval for the population mean in large samples

3. Hence, if $z_{1-\alpha/2}$ and $z_{\alpha/2}$ are the $(1-\alpha/2)$ and $(\alpha/2)$ upper quantiles of the $N(0,1)$, we have

$$\boxed{P(z_{1-\alpha/2} < Z < z_{\alpha/2}) = 1 - \alpha}$$

Standard normal density




4. Therefore $P(\overbrace{z_{1-\alpha/2}}^{-z_{\alpha/2}} < \overbrace{\frac{\bar{X} - \mu_X}{\hat{\sigma}_X/\sqrt{n}}}^{Z} < z_{\alpha/2}) = 1 - \alpha$

# Confidence interval for the population mean in large samples

5. Solve the double inequality for $\mu_X$:

$$-z_{\alpha/2} < \frac{\bar{X} - \mu_X}{\hat{\sigma}_X/\sqrt{n}} < z_{\alpha/2}$$

to obtain the confidence interval estimator

$$(\overbrace{\bar{X} - z_{\alpha/2}\frac{\hat{\sigma}_X}{\sqrt{n}}}^{T_1(\underline{X}_n)}, \overbrace{\bar{X} + z_{\alpha/2}\frac{\hat{\sigma}_X}{\sqrt{n}}}^{T_2(\underline{X}_n)})$$

6. The confidence interval is:

$$\text{CI}_{1-\alpha}(\mu_X) = (\bar{x} - z_{\alpha/2}\frac{\hat{\sigma}_x}{\sqrt{n}}, \bar{x} + z_{\alpha/2}\frac{\hat{\sigma}_x}{\sqrt{n}})$$

# Confidence interval for the population proportion in large samples

## Application of CIs for the population mean in large samples

Let $\underline{X}_n$, $n \geq 30$ be a SRS from a Bernoulli distr. with parameter $p_X$ ($\mu_X = \text{E}[X] = p_X$ and $\sigma_X = \sqrt{p_X(1-p_X)}$). The sample proportion $\hat{p}_X$ is a special case of the sample mean of zero-one observations, $\hat{p}_X = \bar{X}$.

Thus, from the CLT

$$\underbrace{\frac{\hat{p}_X - p_X}{\sqrt{p_X(1-p_X)/n}}}_{\sigma_X/\sqrt{n}} \sim^{\text{approx.}} N(0,1)$$

This result remains true if we use an estimate for the population standard deviation

$$\frac{\hat{p}_X - p_X}{\underbrace{\sqrt{\hat{p}_X(1-\hat{p}_X)}/\sqrt{n}}_{\hat{\sigma}_X/\sqrt{n}}} \sim^{\text{approx.}} N(0,1)$$

Thus, in large samples, the confidence interval for $p_X$ is:

$$\text{CI}_{1-\alpha}(p_X) = \left(\hat{p}_x - z_{\alpha/2}\sqrt{\frac{\hat{p}_x(1-\hat{p}_x)}{n}}, \hat{p}_x + z_{\alpha/2}\sqrt{\frac{\hat{p}_x(1-\hat{p}_x)}{n}}\right)$$

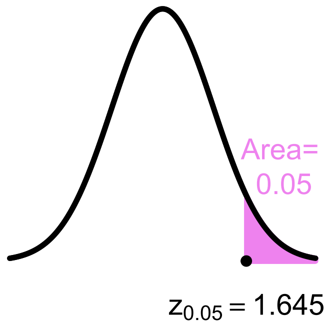
# Example: finding a confidence interval for $p_X$

**Example:** 8.6 (Newbold) A random sample of 344 industrial buyers were asked: "What is your firm's policy for purchasing personnel to follow on accepting gifts from vendors?". For 83 of these buyers, the policy of the firm was for the buyer to make his/her own decision. Find a 90% confidence interval for the population proportion of all buyers who are allowed to make their own decisions.

Population:
$X = 1$ if a buyer makes their own decision and 0 otherwise
$X \sim Bernoulli(p_X)$

SRS: $n = 344$ large

Sample: $\hat{p}_x = \frac{83}{344} = 0.241$




Objective: $CI_{0.9}(p_X) = \left(\hat{p}_X \mp z_{\alpha/2}\sqrt{\frac{\hat{p}_X(1-\hat{p}_X)}{n}}\right)$

$$\hat{p}_X = 0.241 \qquad n = 344$$

$$1-\alpha = 0.9 \Rightarrow \alpha/2 = 0.05$$

$$z_{\alpha/2} = z_{0.05} = 1.645$$

$$CI_{0.9}(p_X) = \left(0.241 \mp 1.645\sqrt{\frac{0.241(1-0.241)}{344}}\right) = (0.241 \mp 0.038) = (0.203, 0.279)$$

Interpretation: We can be 90% confident that the proportion of buyers who make their own decision, $p_X$, falls in $(0.203, 0.279)$

# Confidence interval for the population mean, normal population with unknown variance

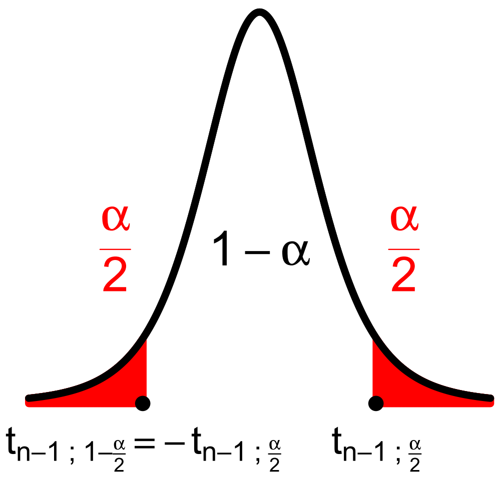
1. Let $\underline{X}_n$ be a SRS of size $n$ from $X$. Under the assumptions:
   - $X$ follows a normal distribution with parameters $\mu_X$ and $\sigma_X^2$
   - $\sigma_X^2$ is unknown (quite realistic)
2. The pivotal quantity for $\mu_X$ is

$$\frac{\bar{X} - \mu_X}{s_X/\sqrt{n}} \sim t_{n-1}$$

# Confidence interval for the population mean, normal population with unknown variance

3. Hence, if $t_{n-1;1-\alpha/2}$ and $t_{n-1;\alpha/2}$ are the $(1-\alpha/2)$ and $(\alpha/2)$ upper quantiles of the $t$ distribution with $n-1$ degrees of freedom (df), we have

$$P(t_{n-1;1-\alpha/2} < \overbrace{T}^{\sim t_{n-1}} < t_{n-1;\alpha/2}) = 1-\alpha$$




$t$ (Student) density

Recall: if $T \sim t_n$, $\mathrm{E}[T] = 0$, $\mathrm{V}[T] = \frac{n}{n-2}$

4. Therefore $P(\overbrace{t_{n-1;1-\alpha/2}}^{-t_{n-1;\alpha/2}} < \overbrace{\frac{\bar{X}-\mu_X}{s_X/\sqrt{n}}}^{T \sim t_{n-1}} < t_{n-1;\alpha/2}) = 1-\alpha$

# Confidence interval for the population mean, normal population with known variance

5. Solve the double inequality for $\mu_X$:

$$-t_{n-1;\alpha/2} < \frac{\bar{X}-\mu_X}{s_X/\sqrt{n}} < t_{n-1;\alpha/2}$$

to obtain the confidence interval estimator

$$(\overbrace{\bar{X} - t_{n-1;\alpha/2}\frac{s_X}{\sqrt{n}}}^{T_1(\underline{X}_n)}, \overbrace{\bar{X} + t_{n-1;\alpha/2}\frac{s_X}{\sqrt{n}}}^{T_2(\underline{X}_n)})$$

6. The confidence interval is:

$$\mathrm{CI}_{1-\alpha}(\mu_X) = (\bar{x} - t_{n-1;\alpha/2}\frac{s_x}{\sqrt{n}}, \bar{x} + t_{n-1;\alpha/2}\frac{s_x}{\sqrt{n}})$$

# Example: finding a confidence interval for $\mu_X$

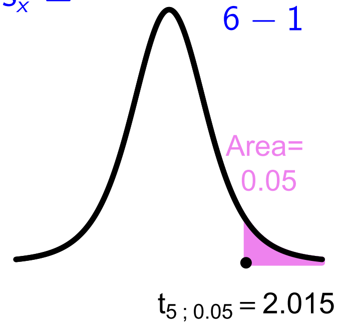
**Example:** 8.4 (Newbold) A random sample of six cars from a particular model year had the following fuel consumption figures, in mpg: 18.6, 18.4, 19.2, 20.8, 19.4, 20.5. Find a 90% confidence interval for the population mean fuel consumption, assuming that the population distribution is normal.

Population:
$X$ = "mpg of a car from the model year" $X \sim N(\mu_X, \sigma_X^2)$ $\sigma_X^2$ unknown

SRS: $n = 6$ small

Sample: $\bar{x} = \frac{116.9}{6} = 19.4833$

$$s_x^2 = \frac{2282.41 - 6(19.4833)^2}{6-1} = 0.96$$




Objective: $CI_{0.9}(\mu_X) = \left(\bar{x} \mp t_{n-1;\alpha/2} \frac{s_x}{\sqrt{n}}\right)$

$$s_x = \sqrt{0.96} = 0.98$$

$$n = 6 \qquad \bar{x} = 19.48$$

$$1 - \alpha = 0.9 \Rightarrow \alpha/2 = 0.05$$

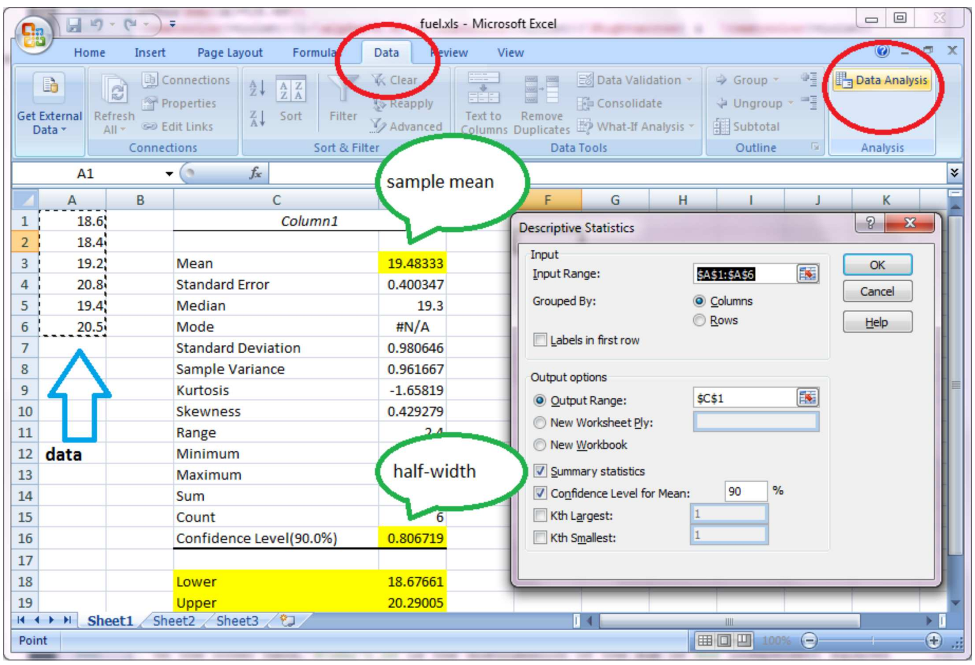
$$t_{n-1;\alpha/2} = t_{5;0.05} = 2.015$$

$$CI_{0.9}(\mu_X) = \left(19.48 \mp 2.105 \frac{0.98}{\sqrt{6}}\right) = (19.48 \mp 0.81) = (18.67, 20.29)$$

Interpretation: We can be 90% confident that the population mean fuel consumption for these cars, $\mu_X$, is between 18.67 and 20.29

# Example: finding a confidence interval for $\mu_X$

**Example:** 8.4 (cont.) in Excel: Go to menu: Data, submenu: Data Analysis, choose function: Descriptive Statistics.
Column A (data), in yellow (sample mean, half-width $t_{n-1;\alpha/2} \frac{s_x}{\sqrt{n}}$, lower end-point (cell D3-D16), upper end-point (cell D3+D16)).



# $t$ and $\chi^2$ distributions

- Recall that $T \sim t_n$ if $T = \frac{Z}{\sqrt{\chi^2_n/n}}$, where $Z \sim N(0,1)$ and $\chi^2_n$ follows a chi-square distribution with df $= n$, independent of $Z$.
- On the other hand, $\chi^2_n$ is the distribution of the sum of $n$ independent squared $N(0,1)$ random variables.
- Note that the rescaled sample quasi variance follows a chi-square distribution with $n-1$ degrees of freedom

$$\frac{(n-1)s_X^2}{\sigma_X^2} = \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{\sigma_X^2} = \sum_{i=1}^n \left(\frac{X_i - \bar{X}}{\sigma_X}\right)^2 \sim \chi^2_{n-1}$$

Why $n-1$ and not $n$?

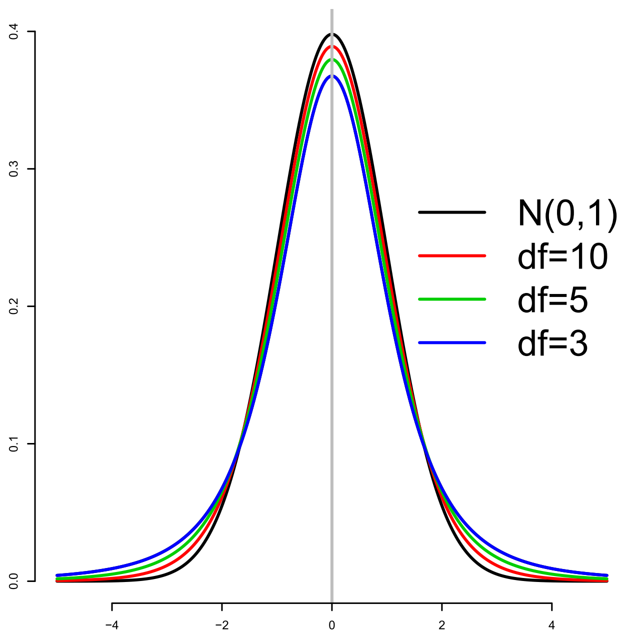
If we knew $\mu_X$, the number of degrees of freedom would be $n$, because we would have $n$ *iid* random variables $\frac{X_i - \mu_X}{\sigma_X}$

Since we have to estimate $\mu_X$ with $\bar{X}$, the df are $n-1$, because we only have $n-1$ *iid* random variables $\frac{X_i - \bar{X}}{\sigma_X}$ (once you know $n-1$ of them, you can figure out the remaining one)

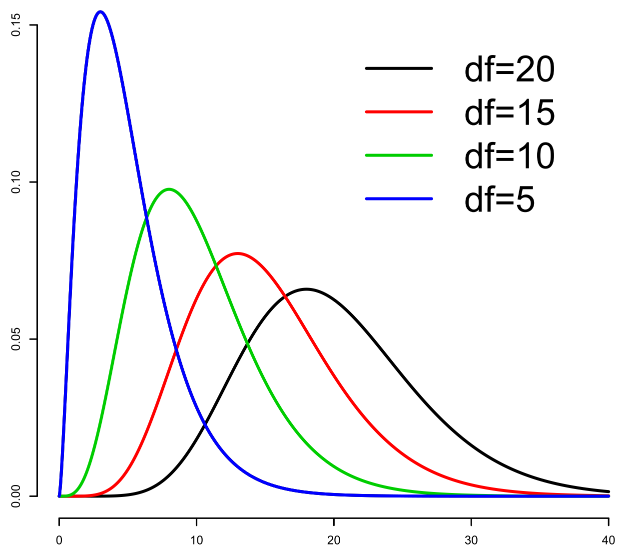
We say that one degree of freedom is used up to estimate $\mu_X$

# $t$ and $\chi^2$ distributions

$t$ and $N(0,1)$ densities




$\chi^2$ densities



# Confidence interval for the population variance, normal population

1. Let $\underline{X}_n$ be a SRS of size $n$ from $X$. Under the assumptions:
   - $X$ follows a normal distribution with parameter $\sigma_X^2$
2. The pivotal quantity for $\sigma_X^2$ is

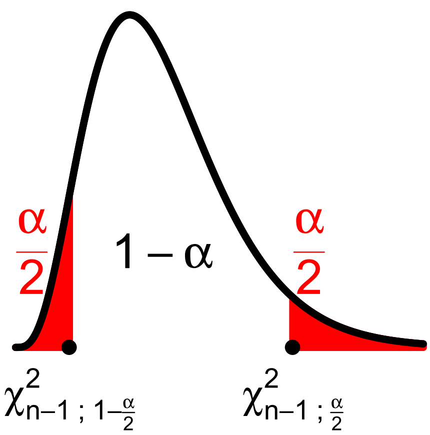
$$\boxed{\frac{(n-1)s_X^2}{\sigma_X^2} \sim \chi^2_{n-1}}$$

# Confidence interval for the population variance, normal population

3. Hence, if $\chi^2_{n-1;1-\alpha/2}$ and $\chi^2_{n-1;1-\alpha/2}$ are the $(1-\alpha/2)$ and $(\alpha/2)$ upper quantiles of the chi-square distribution with $n-1$ degrees of freedom, we have

$$\boxed{P(\chi^2_{n-1;1-\alpha/2} < \chi^2_{n-1} < \chi^2_{n-1;\alpha/2}) = 1-\alpha}$$

Chi-square density




Recall: $\mathrm{E}[\chi^2_n] = n$, $\mathrm{V}[\chi^2_n] = 2n$

4. Therefore $P(\chi^2_{n-1;1-\alpha/2} < \overbrace{\frac{(n-1)s_X^2}{\sigma_X^2}}^{\chi^2_{n-1}} < \chi^2_{n-1;\alpha/2}) = 1-\alpha$

# Confidence interval for the population variance, normal population

5. Solve the double inequality for $\sigma_X^2$:

$$\begin{array}{ccccc} \chi^2_{n-1;1-\alpha/2} & < & \frac{(n-1)s_X^2}{\sigma_X^2} & < & \chi^2_{n-1;\alpha/2} \\ \frac{1}{\chi^2_{n-1;1-\alpha/2}} & > & \frac{\sigma_X^2}{(n-1)s_X^2} & > & \frac{1}{\chi^2_{n-1;\alpha/2}} \\ \frac{(n-1)s_X^2}{\chi^2_{n-1;1-\alpha/2}} & > & \sigma_X^2 & > & \frac{(n-1)s_X^2}{\chi^2_{n-1;\alpha/2}} \end{array}$$

to obtain the confidence interval estimator

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

6. The confidence interval is:

$$\text{CI}_{1-\alpha}(\sigma_X^2) = \left( \frac{(n-1)s_x^2}{\chi^2_{n-1;\alpha/2}}, \frac{(n-1)s_x^2}{\chi^2_{n-1;1-\alpha/2}} \right)$$

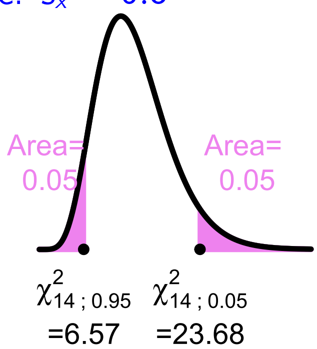
# Example: finding a confidence interval for $\sigma_X^2$ and $\sigma_X$

**Example:** 8.8 (Newbold) A random sample of fifteen pills for headache relief showed a quasi standard deviation of 0.8% in the concentration of the active ingredient. Find a 90% confidence interval for the population variance for these pills. How would you obtain a CI for the population standard deviation?

Population:
$X$ = "concentration of an active ingredient in a pill (in %)"
$X \sim N(\mu_X, \sigma_X^2)$

SRS: $n = 15$

Sample: $s_x = 0.8$




Objective: $CI_{0.9}(\sigma_X^2) = \left( \frac{(n-1)s_X^2}{\chi^2_{n-1;\alpha/2}}, \frac{(n-1)s_X^2}{\chi^2_{n-1;1-\alpha/2}} \right)$

$s_x^2 = 0.8^2 = 0.64 \qquad n = 15$

$$\begin{aligned} 1-\alpha = 0.9 &\Rightarrow \alpha/2 = 0.05 \\ \chi^2_{n-1;1-\alpha/2} &= \chi^2_{14;0.95} = 6.57 \\ \chi^2_{n-1;\alpha/2} &= \chi^2_{14;0.05} = 23.68 \\ CI_{0.9}(\sigma_X^2) &= \left( \frac{14(0.64)}{23.68}, \frac{14(0.64)}{6.57} \right) \\ &= (0.378, 1.364) \Rightarrow \\ CI_{0.9}(\sigma_X) &= (\sqrt{0.378}, \sqrt{1.364}) \\ &= (0.61, 1.17) \end{aligned}$$

To obtain $CI(\sigma_X)$ we apply $\sqrt{}$ to the end-points of $CI(\sigma_X^2)$

# Confidence intervals formulae

## Summary for one population

- Let $\underline{X}_n$ be a simple random sample from a population $X$ with mean $\mu_X$ and variance $\sigma_X^2$

| Parameter | Assumptions | Pivotal quantity | $(1-\alpha)$ Conf. Interval |
|---|---|---|---|
| Mean | Normal data<br>Known variance | $\frac{\bar{X}-\mu_X}{\sigma_X/\sqrt{n}} \sim N(0,1)$ | $\mu_X \in \left(\bar{x} - z_{\alpha/2}\frac{\sigma_X}{\sqrt{n}}, \bar{x} + z_{\alpha/2}\frac{\sigma_X}{\sqrt{n}}\right)$ |
| | Nonnormal data<br>Large sample | $\frac{\bar{X}-\mu_X}{\hat{\sigma}_X/\sqrt{n}} \sim_{approx.} N(0,1)$ | $\mu_X \in \left(\bar{x} - z_{\alpha/2}\frac{\hat{\sigma}_X}{\sqrt{n}}, \bar{x} + z_{\alpha/2}\frac{\hat{\sigma}_X}{\sqrt{n}}\right]$ |
| | Bernoulli data<br>Large sample | $\frac{\hat{p}_X-p_X}{\sqrt{\hat{p}_X(1-\hat{p}_X)/n}} \sim_{approx.} N(0,1)$ | $p_X \in \left(\hat{p}_x \mp z_{\alpha/2}\sqrt{\frac{\hat{p}_x(1-\hat{p}_x)}{n}}\right]$ |
| | Normal data<br>Unknown variance | $\frac{\bar{X}-\mu_X}{s_X/\sqrt{n}} \sim t_{n-1}$ | $\mu_X \in \left(\bar{x} - t_{n-1,\alpha/2}\frac{s_X}{\sqrt{n}}, \bar{x} + t_{n-1,\alpha/2}\frac{s_X}{\sqrt{n}}\right)$ |
| Variance | Normal data | $\frac{(n-1)s_X^2}{\sigma_X^2} \sim \chi^2_{n-1}$ | $\sigma_X^2 \in \left(\frac{(n-1)s_X^2}{\chi^2_{n-1;\alpha/2}}, \frac{(n-1)s_X^2}{\chi^2_{n-1;1-\alpha/2}}\right)$ |
| Standard dev. | Normal data | $\frac{(n-1)s_X^2}{\sigma_X^2} \sim \chi^2_{n-1}$ | $\sigma_X \in \left(\sqrt{\frac{(n-1)s_X^2}{\chi^2_{n-1;\alpha/2}}}, \sqrt{\frac{(n-1)s_X^2}{\chi^2_{n-1;1-\alpha/2}}}\right)$ |

# Confidence intervals for the population mean: when to use what?