

# 7 Statistical Intervals Based on a Single Sample

## 7.1 Basic Properties of Confidence Intervals

**Unrealistic scenario:** Let  $X_1, \dots, X_n$  be a **simple random sample** from a  $N(\mu, \sigma)$  distribution where  $\mu$  is unknown but  $\sigma$  is known.

Why is this scenario “unrealistic”?


What is a reasonable **point estimator** of  $\mu$ ?

What is the **mean** of  $\bar{X}$ ?

What is the **precision** of  $\bar{X}$ ?

**Example:** Find the  $z$ -score such that  $P(-z < Z < z) = 0.95$ , where  $Z \sim N(0, 1)$ .





$t$ -table, p. 671

$\nu$	$\alpha$						
	.10	.05	.025	.01	.005	.001	.0005
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
50	1.299	1.676	2.009	2.403	2.678	3.261	3.496
60	1.296	1.671	2.000	2.390	2.660	3.232	3.460
120	1.289	1.658	1.980	2.358	2.617	3.160	3.373
$\infty$	1.282	1.645	1.960	2.326	2.576	3.090	3.291

□

**Example:** Derive a 95% confidence interval on  $\mu$ .

□

**Example:** (unrealistic) Let  $X_1, \dots, X_6$  be a simple random sample from a  $N(\mu, \sigma = 10)$  distribution. The observations are  $(X_1, \dots, X_6) = (56, 45, 30, 61, 38, 52)$ .

(a) Construct a 95% confidence interval on  $\mu$ .

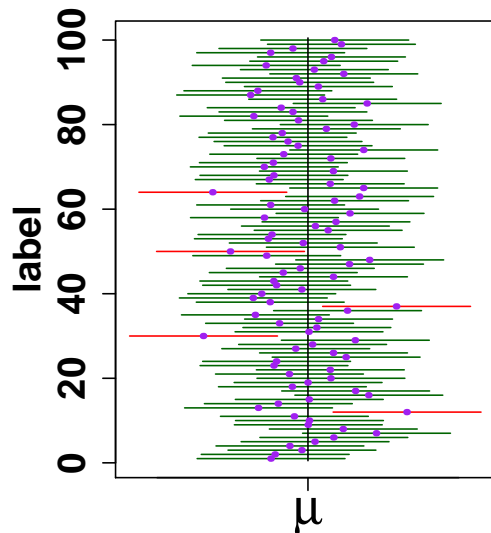
(b) State the **layman's** interpretation of the confidence interval.

We are 95% **confident** that the population mean,  $\mu$ , lies between 38.998 and 55.002.

## Interpreting a Confidence Interval

**Mathematically rigorous interpretation of the confidence interval:** If we repeat the sampling procedure many times to construct many 95% confidence intervals on  $\mu$ , the population mean, then approximately 95% of these 95% confidence intervals will contain the true value of  $\mu$ .

**95% of these C.I.s contain  $\mu$**

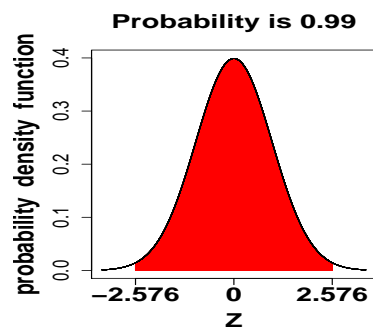


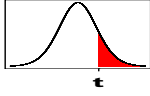
**CONFIDENCE INTERVALS**

## Other Levels of Confidence

**Example:** (Revisit; unrealistic) Let  $X_1, \dots, X_6$  be a **simple random sample** from a  $N(\mu, \sigma = 10)$  distribution. The observations are  $(X_1, \dots, X_6) = (56, 45, 30, 61, 38, 52)$ .

(a) Construct a **99%** confidence interval on  $\mu$ .





$t$ -table, p. 671

$\nu$	$\alpha$						
	.10	.05	.025	.01	.005	.001	.0005
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
50	1.299	1.676	2.009	2.403	2.678	3.261	3.496
60	1.296	1.671	2.000	2.390	2.660	3.232	3.460
120	1.289	1.658	1.980	2.358	2.617	3.160	3.373
$\infty$	1.282	1.645	1.960	2.326	2.576	3.090	3.291

(b) State the **layman's** interpretation of the confidence interval.

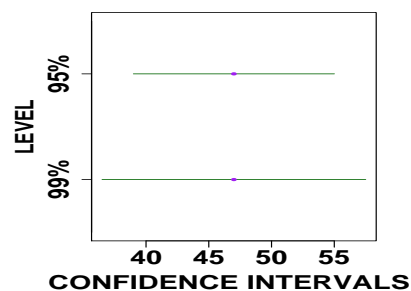
We are 99% **confident** that the population mean,  $\mu$ , lies between 36.5 and 57.5.

(c) State the **mathematically rigorous** interpretation of the confidence interval.

If we repeat the sampling procedure many times to construct many 99% confidence intervals on  $\mu$ , the population mean, then approximately 99% of these 99% confidence intervals will contain the true value of  $\mu$ .

(d) Is the following interpretation correct:  $P(36.5 < \mu < 57.5) = 99\%$ ?

(e) Which is wider, the 95% confidence interval or the 99% confidence interval?



(f) How can we increase the **level** of confidence without increasing the **width** of confidence interval?

□

**Example:** Do you prefer a **high** level (e.g., 99.9% level) of confidence or a **low** level (e.g., 50% level) of confidence in the following? You work for a bomb squad. A *red* wire and a *blue* wire are remaining. Cutting the *correct* wire results in *life*, but cutting the *wrong* wire results in *death*. Your partner says, “Cut the *red* wire.” You respond, “How confident are you?”

□

**Example:** Do you prefer a **wide** confidence interval or a **narrow** confidence interval in the following?

The Joint United Nations Programme on HIV/AIDS (UNAIDS) is 95% confident that the population proportion of adults aged 15 to 49 from Botswana who are infected with HIV is between 23% and 32%.

I am almost 100% confident that the population proportion of adults aged 15 to 49 from Botswana who are infected with HIV is between 0.001% and 99.999%.

□

What is the optimal confidence level; e.g., 90%, 95%, or 99%?

## Sample size determination

The **wider** the confidence interval, the **less** precise the confidence interval is.

The **width** of a confidence interval can be decreased by

(a) decreasing the level of confidence,

(b) decreasing  $\sigma$ ,

- (c) increasing  $n$ .

When taking a simple random sample from a  $N(\mu, \sigma)$  distribution where  $\sigma$  is known but  $\mu$  is unknown:

- (a) A **confidence interval** on  $\mu$  is  $\bar{X} \pm z\sigma/\sqrt{n}$ .
- (b) The **standard error** on  $\bar{X}$  is  $\sigma_{\bar{x}} = \sigma/\sqrt{n}$ .
- (c) The **error margin** or **margin of error** on  $\bar{X}$  is  $z \sigma_{\bar{x}} = z \sigma/\sqrt{n}$ , which is **half the width** of the confidence interval.
- (d) The **error of estimation** is  $|\bar{X} - \mu|$ .
- (e) The **bound on the error of estimation** is the **error margin**; i.e., with 95% confidence we want  $|\bar{X} - \mu| \leq z \sigma/\sqrt{n}$ .

**Example:** (unrealistic) Let  $X_1, \dots, X_n$  be a **simple random sample** from a  $N(\mu, \sigma)$  distribution where  $\mu$  is unknown but  $\sigma$  is known. Determine the **sample size** needed for a specific width  $w$  and specific level of confidence.

**Example:** (Revisit; unrealistic) Let  $X_1, \dots, X_6$  be a **simple random sample** from a  $N(\mu, \sigma = 10)$  distribution. The observations are  $(X_1, \dots, X_6) = (56, 45, 30, 61, 38, 52)$ . Determine the **sample size** needed for a 95% confidence interval with width 3 (i.e., with bound on the error estimation equal to 1.5).

□

## 7.2 Large-Sample Confidence Intervals for a Population Mean and Proportion

### Population Mean

*Realistic scenario:* Let  $X_1, \dots, X_n$  be a **simple random sample** from a population with unknown mean  $\mu$  and unknown positive finite standard deviation  $\sigma$ .

If  $n$  is large (usually  $n \geq 30$ , if neither tail of the distribution is too heavy), then

$$\bar{X} \stackrel{\text{approx.}}{\sim} N(\mu, \sigma/\sqrt{n})$$

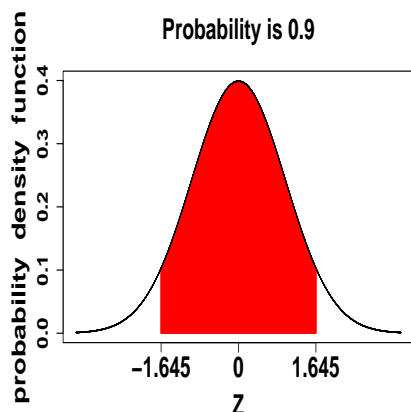
by the Central Limit Theorem, and an approximate  $100(1 - \alpha)\%$  level confidence interval on  $\mu$  is

$$\bar{X} \pm z_{\alpha/2} s/\sqrt{n}.$$

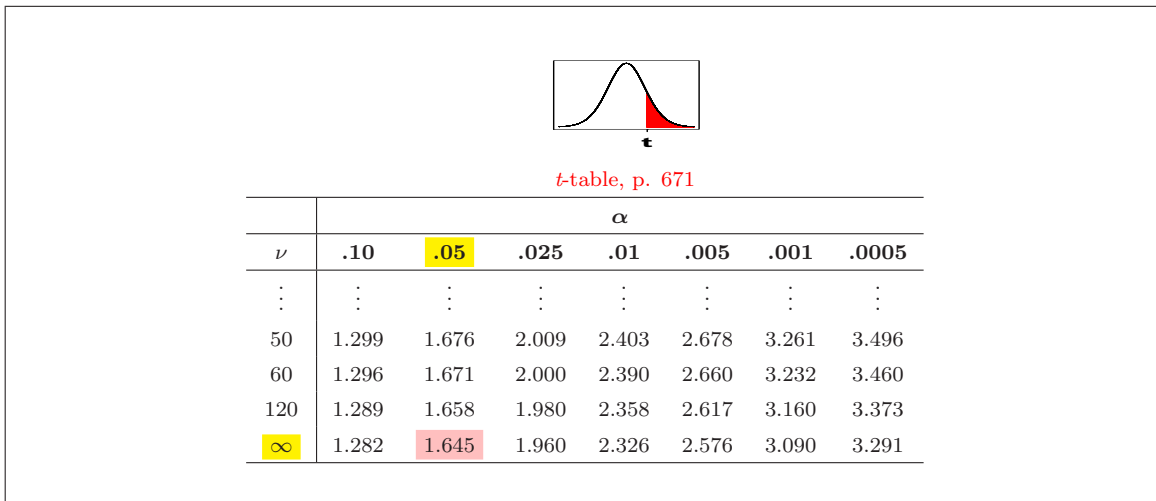
**Remark:** The above confidence interval is valid for **any** distribution with positive finite standard deviation.

**Remark:** If the distribution has at least one heavy tail, then  $n$  might need to be very large in order for this approximation (on the confidence interval) to be reasonable.

**Example:** Suppose 100 observations are sampled independently from a population with unknown mean  $\mu$  and unknown positive finite standard deviation  $\sigma$ . Suppose the sample mean  $\bar{X} = 83$  and the sample standard deviation  $s = 13$ , and further suppose that no outliers are in the data. Construct a **90%** confidence interval on  $\mu$ .







## Population Proportion

Let  $X \sim \text{Binomial}(n, p)$ .

Recall for section 4.3:  $X \stackrel{\text{approx.}}{\sim} N(\mu_x = np, \sigma_x = \sqrt{np(1-p)})$  if  $np \geq 10$  and  $n(1-p) \geq 10$ .

Determine the approximate distribution of  $\hat{p}$ , the sample proportion, if  $np \geq 10$  and  $n(1-p) \geq 10$ .

**Example:** Derive a large-sample 95% confidence interval on  $p$ .

□

A more accurate approximation to the confidence interval is based on completing the square, equation (7.10).

For a sample proportion,  $\hat{p}$  (based on a simple random sample):

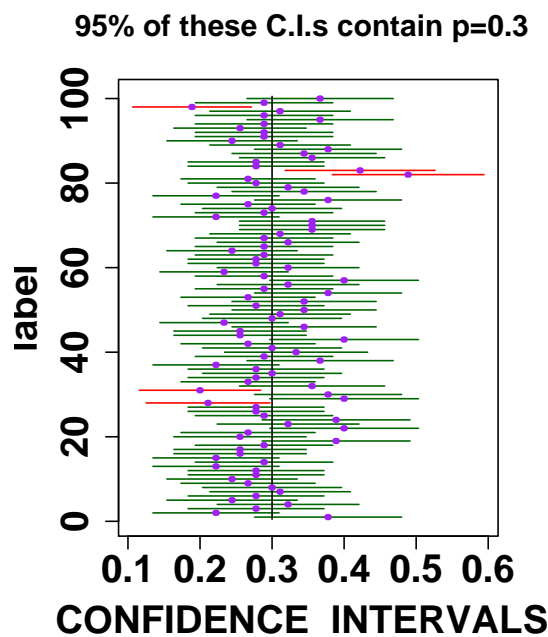
(a) (standard error) =  $\sigma_{\hat{p}} = \sqrt{p(1-p)/n} \approx \sqrt{\hat{p}(1-\hat{p})/n}$ .

(b) For 95% confidence, the (margin of error)  
 =  $z \times$  (standard error)  $\approx 1.96\sqrt{\hat{p}(1-\hat{p})/n}$ .

- (c) For  $n\hat{p} \geq 10$  and  $n(1 - \hat{p}) \geq 10$ , the 95% confidence interval on *unknown, fixed*  $p$  is  $\hat{p} \pm (\text{margin of error}) = \hat{p} \pm 1.96\sqrt{\hat{p}(1 - \hat{p})/n}$ .

**Layman’s interpretation:** We are 95% confident that the population proportion,  $p$ , lies in the confidence interval.

**Mathematically rigorous interpretation:** If we repeat the sampling procedure many times to construct many 95% confidence intervals on  $p$ , then approximately 95% of these 95% confidence intervals will contain the true value of  $p$ .



**Example:** *Estimating the success rate at the Charlottesville fertility clinic, called University of Virginia Assisted Reproductive Technology (ART) program.*

64 women no older than 40 years-old attempted to get pregnant from using services at the UVa clinic.

Do these 64 women represent a simple random sample of women from the U.S.?

The population consists of all women no older than 40, from similar regions, who

would seek clinical pregnancy services from this type of clinic.

Among those 64 women, 20 successfully gave live births (i.e., no miscarriages).

We want to estimate  $p$ , the population proportion of *similar* women who would give live births when using this clinic.

Hence,  $p$  is the population success rate of this clinic.

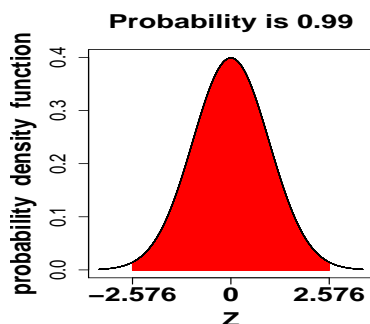
$X = 20$  the number of women who successfully gave live births.

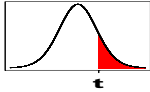
- (a) Determine the appropriate *point estimate* of  $p$ , the population success rate of this clinic.
- (b) Construct a **95%** confidence interval on  $p$ , the population success rate of this clinic.

**Layman’s interpretation:** We are 95% confident that the population success rate of this clinic lies between 0.199 and 0.426.

**Mathematically rigorous interpretation:** If we repeat the sampling procedure many times to construct many 95% confidence intervals on  $p$ , the population success rate of this clinic, then approximately 95% of these 95% confidence intervals will contain  $p$ .

- (c) Which of the following are correct interpretations?
  - $P(0.199 < p < 0.426) \approx 0.95$
  - $P(0.199 < \hat{p} < 0.426) \approx 0.95$
- (d) Now suppose that we want a **99%** confidence interval on  $p$ .

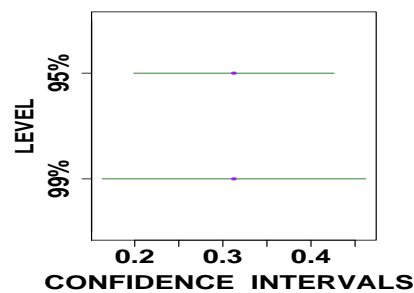




t-table, p. 671

$\nu$	$\alpha$						
	.10	.05	.025	.01	.005	.001	.0005
∞	1.282	1.645	1.960	2.326	2.576	3.090	3.291

(e) Which confidence interval is wider?



(f) How can we increase the **level** of confidence without increasing the **width** of confidence interval?

□

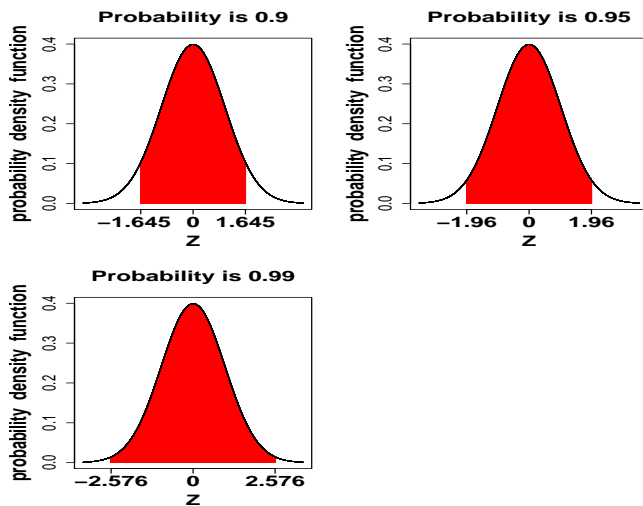
### Sample size determination, when estimating $p$

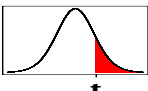
Recall: For  $n\hat{p} \geq 10$  and  $n(1 - \hat{p}) \geq 10$ , a **confidence interval on  $p$** , the unknown population proportion, is

$$\hat{p} \pm z \sqrt{\frac{\hat{p}(1 - \hat{p})}{n}}$$

The **margin of error** on  $\hat{p}$  is  $z\sqrt{\hat{p}(1 - \hat{p})/n}$ , which is **half the width** of the confidence interval.

Suppose we want to construct a 95% confidence interval on  $p$ , where the **width**,  $w$ , is selected prior to drawing the sample.





$t$

*t*-table, p. 671

	$\alpha$						
$\nu$	.10	.05	.025	.01	.005	.001	.0005
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
50	1.299	1.676	2.009	2.403	2.678	3.261	3.496
60	1.296	1.671	2.000	2.390	2.660	3.232	3.460
120	1.289	1.658	1.980	2.358	2.617	3.160	3.373
$\infty$	1.282	1.645	1.960	2.326	2.576	3.090	3.291

What sample size,  $n$ , is needed?

Solve for  $n$  in

$$w/2 = 1.96 \sqrt{\hat{p}(1 - \hat{p})/n}$$

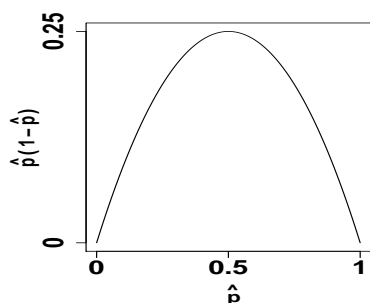
to obtain

$$n = \hat{p}(1 - \hat{p})(2 \times 1.96/w)^2.$$

What is the drawback when using the above formula for  $n$ ?

Two options:

- (a) Use a preliminary *point estimate*  $\hat{p}$ , and then compute
 
$$n = 4\hat{p}(1 - \hat{p})(1.96/w)^2, \quad \text{OR}$$
- (b) The maximum value of  $n = \hat{p}(1 - \hat{p})(2 \times 1.96/w)^2$  occurs when  $\hat{p} = 0.5$ , so use
 
$$n = (1.96/w)^2 \quad (\text{conservative sample size}).$$



**Example:** Revisit the Charlottesville fertility clinic. A sample of 64 women resulted in 20 live births. However, the population success rate,  $p$ , of this clinic is unknown. What sample size  $n$  is needed to obtain a **95%** confidence interval on  $p$  with **width** approximately equal to **0.12**, using:

- (a) 0.3125, as the initial *point estimate* of  $p$ ?
- (b) no initial *point estimate* of  $p$ ?

**Example:** Revisit the Charlottesville fertility clinic, again! What sample size  $n$  is needed to obtain a **90%** confidence interval on  $p$  with **margin of error** approximately equal to **0.06**, using:

- (a) 0.3125, as the initial *point estimate* of  $p$ ?

- (b) no initial *point estimate* of  $p$ ?
- (c) Repeat part (b) using  $w = 0.06$ .

## 7.3 Intervals Based on a Normal Population Distribution

### The $t$ distribution

Let  $X_1, \dots, X_n$  be independent observations from a population with mean  $\mu$  and positive finite standard deviation  $\sigma$ .

- (a)  $\mu_{\bar{X}} = \mu$
- (b)  $\sigma_{\bar{X}} = \sigma/\sqrt{n}$  (called the **standard error** of  $\bar{X}$ )
- (c) (A version of the Central Limit Theorem) The sample mean,  $\bar{X}$ , is approximately normally distributed for **large**  $n$  (usually  $n \geq 30$ , if neither tail of the distribution is too heavy).
- (d) (A special case) The sample mean,  $\bar{X}$ , is approximately normally distributed (for **any** sample size  $n$ ), if the **original population** is approximately **normally distributed**.

Therefore, for independent (or nearly independent observations) and positive finite  $\sigma$ , if the **original population is approximately normal OR  $n$  is large**, then

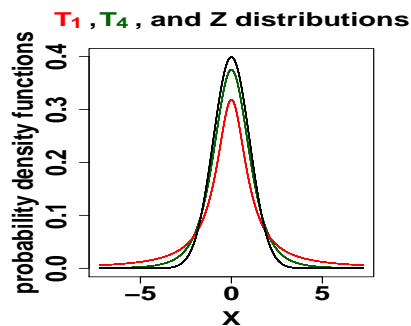
$$Z = \frac{\bar{X} - \mu_{\bar{X}}}{\sigma_{\bar{X}}} = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \underset{\text{approx.}}{\sim} N(0, 1), \text{ and}$$
$$T = \frac{\bar{X} - \mu}{s/\sqrt{n}} \underset{\text{approx.}}{\sim} t_{n-1}$$

Thus,  $T$  has a  $t$  distribution with  $(n - 1)$  degrees of freedom.

The  $t$  distribution is symmetric about zero, has no units, and has heavier tails than the standard normal distribution.

As the number of degrees of freedom gets large, then  $s$  “converges” to  $\sigma$ , so the  $t$  distribution starts to converge to the standard normal distribution.

**Example:** Below are the *probability density functions* of a  $t$  distribution with *one* degree of freedom, a  $t$  distribution with *four* degrees of freedom, and *standard normal distribution*.



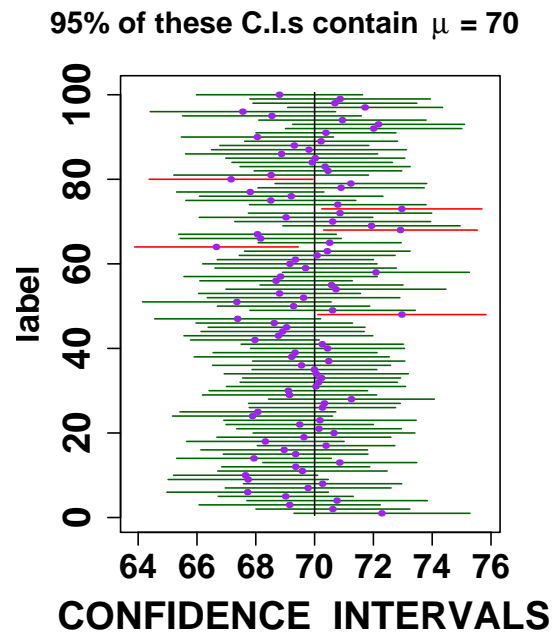
□

### Confidence interval on $\mu$

For independent (or nearly independent observations) and positive finite  $\sigma$ , if **the original population is approximately normal OR  $n$  is large**, then a confidence interval on  $\mu$  is

$$\bar{X} \pm (\text{margin of error}) = \bar{X} \pm t_{n-1}(\text{standard error}) = \bar{X} \pm t_{n-1}s/\sqrt{n}.$$

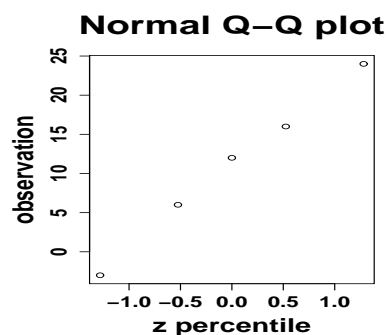




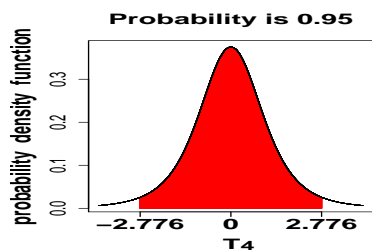
**Example:** A sample of individuals participating in a rigorous exercise program results in the following weight losses in pounds: {16, 6, 24, -3, 12}.

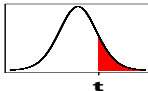
The population consists of all *similar* individuals who would be willing to participate in this rigorous exercise program, if offered the opportunity.

(a) Are the assumptions for constructing a confidence interval satisfied?



(b) Construct a 95% confidence interval on the population mean weight loss.





$t$

*t*-table, p. 725

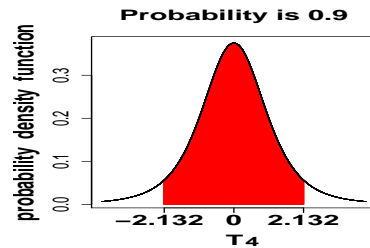
$\nu$	$\alpha$						
	.10	.05	.025	.01	.005	.001	.0005
1	3.078	6.314	12.706	31.821	63.657	318.309	636.619
2	1.886	2.920	4.303	6.965	9.925	22.327	31.599
3	1.638	2.353	3.182	4.541	5.841	10.215	12.924
4	1.533	2.132	2.776	3.747	4.604	7.173	8.610
5	1.476	2.015	2.571	3.365	4.032	5.893	6.869
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$

**Layman’s interpretation:** We are 95% confident that the population mean weight loss,  $\mu$ , of this exercise program is between  $-1.66$  pounds and  $23.66$  pounds.

**Mathematically rigorous interpretation:** If we repeat the sampling procedure many times to construct many 95% confidence intervals on  $\mu$ , the population mean weight loss of this exercise program, then approximately 95% of these 95% confidence intervals will contain the true value of  $\mu$ .

- (c) Which of the following are correct interpretations?
- $P(-1.66 \text{ pounds} < \mu < 23.66 \text{ pounds}) \approx 0.95$
  - $P(-1.66 \text{ pounds} < \bar{X} < 23.66 \text{ pounds}) \approx 0.95$
  - 95% of the population of weight losses lies between  $-1.66$  pounds and  $23.66$  pounds.
  - $P(-1.66 \text{ pounds} < X < 23.66 \text{ pounds}) \approx 0.95$

(d) Construct a **90%** confidence interval on the population mean weight loss.



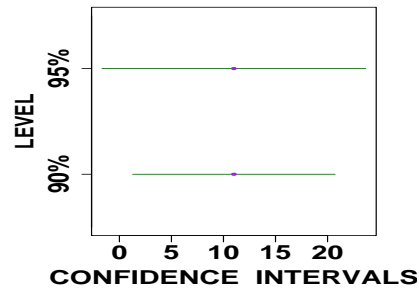
*t*-table, p. 725

$\nu$	$\alpha$						
	.10	.05	.025	.01	.005	.001	.0005
1	3.078	6.314	12.706	31.821	63.657	318.309	636.619
2	1.886	2.920	4.303	6.965	9.925	22.327	31.599
3	1.638	2.353	3.182	4.541	5.841	10.215	12.924
4	1.533	2.132	2.776	3.747	4.604	7.173	8.610
5	1.476	2.015	2.571	3.365	4.032	5.893	6.869
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$

**Layman's interpretation:** We are 90% confident that the population mean weight loss,  $\mu$ , of this exercise program is between 1.28 pounds and 20.72 pounds.

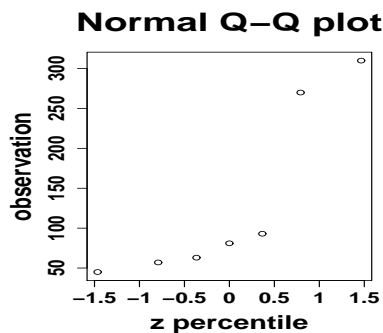
**Mathematically rigorous interpretation:** If we repeat the sampling procedure many times to construct many 90% confidence intervals on  $\mu$ , the population mean weight loss of this exercise program, then approximately 90% of these 90% confidence intervals will contain the true value of  $\mu$ .

(e) Which confidence interval is wider?



□

**Example:** In a simple random sample from a large population, the following observations were taken: {45, 310, 93, 63, 81, 270, 57}. Construct a **95%** confidence interval on the population mean.



□

$(\bar{X} - \mu)/(s/\sqrt{n})$  is approximately  $t$ -distributed, even for **moderate** sample sizes with **moderate** departures from **normality**.

**Robustness:** The  $t$ -confidence interval is **robust**. If the sample size is moderate (e.g., around 20) and the distribution is NOT heavy-tailed, then the  $t$ -confidence interval typically is precise anyway.

## A Prediction Interval for a Single Future Value

*Realistic scenario:* Let  $X_1, \dots, X_n$  be a **simple random sample** from a **normal** population with unknown mean  $\mu$  and unknown positive finite standard deviation  $\sigma$ . Determine an interval for **predicting a new** observation.

*Assume:* The **new** observation is **independent** of the past  $n$  observations.

*Goal:* **Predict** a new observation,  $X_{n+1}$ , which is  $N(\mu, \sigma)$ .

Determine the distribution of  $\bar{X} - X_{n+1}$ .

Determine a  $100(1 - \alpha)$  **prediction interval** on  $X_{n+1}$ .

**Example:** *Revisit.* A sample of individuals participating in a rigorous exercise program results in the following weight losses in pounds:  $\{16, 6, 24, -3, 12\}$ .

The population consists of all *similar* individuals who would be willing to participate in this rigorous exercise program, if offered the opportunity.

For a **new** individual from this population, **predict** the individual's weight loss at a 95% level.

**Interpretation:** Approximately 95% of 95% prediction intervals on  $X_{n+1}$  will contain the random variable  $X_{n+1}$  after many replications of  $(n + 1)$  observations.

□

**Remark:** A **prediction** interval is **wider** than a **confidence** interval.