Sample from the Distribution

Selected Topics

1. Independent events and random variables
2. Sample from the distribution
3. Confidence intervals
4. Hypothesis testing

Independent Events

Two events are independent if the occurrence of one does not in any way affect the probability of the occurrence of the other.

Example: You toss a quarter and a dime. Let A be the event "*Head* on the quarter" and B the event "*Head* on the dime."

Then the probability of occurrence of both A and B is $^1/_4$.

Event "both A and B" is denoted as $A \cap B$. Hence $P(A \cap B) = {}^1/_4$.

Notice that $P(A) = P(B) = {}^1/_2$.

Definition: Events A and B are independent if $P(A \cap B) = P(A) \cdot P(B)$

Conditional Probability

Definition: Conditional probability of event A occurring in an experiment given that event B occurred, denoted by $P(A|B)$ ("probability of A given B"), is

$$P(A|B) = \frac{P(A \cap B)}{P(B)}.$$

Notice that $P(B) > 0$ (think what "given that event B occurred" implies).

Example: A fair die is rolled and you are guessing the outcome. You want to claim the outcome is six. Your weird buddy is on the side and he claims that although he did not see the outcome, he is certain that no *dot* was in the middle. Does this information help you? Quantify it.

Solution: Let S be the event "the outcome is six" and E "the outcome is even."

Your buddy knows that event E occurred in this experiment.

Absent your buddy, your probability of guessing the outcome is $P(S) = {}^1/_6$.

If your buddy whispers to you, your probability of guessing the outcome is

$$P(S|E) = \frac{P(S \cap E)}{P(E)} = \frac{P(S)}{P(E)} = \frac{{}^1/_6}{{}^1/_2} = \frac{1}{3}.$$

Independent Events and Cond. Probability

Note: If events A and B are independent and $P(B) > 0$ then

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{P(A) \cdot P(B)}{P(B)} = P(A).$$

$$\left(\text{and similarly } P(B|A) = \cdots = P(B) \text{ if } P(A) > 0\right)$$

Hence the knowledge of occurrence of event B does not change the probability of occurrence of A.

The converse also holds: if the knowledge of occurrence of event B does not change the probability of occurrence of A, i.e., if

$$P(A|B) = P(A) \quad \text{(note that } P(B) > 0)$$

then the events A and B are independent.

Independence of random variables

Two random variables are independent if the outcome of one variable does not in any way influence the outcome of another.

Example: Roll a die and toss a coin. If X is the outcome of the die (1, 2, 3, 4, 5, or 6) and Y the outcome of the coin (0 or 1), then X and Y are independent random variables.

Definition: Random variables X and Y are independent if

$$P(a < X \le b, c < Y \le d) = P(a < X \le b) \cdot P(c < Y \le d),$$

for all real numbers a, b, c, d such that $a < b$ and $c < d$.

Note: Discrete random variables X and Y are independent if and only if

$$P(X = k,\ Y = m) = P(X = k) \cdot P(Y = m),$$

for all discrete values k that X assumes and all discrete values m that Y assumes.

Sample from the distribution

Random variables $X_1, X_2, \ldots, X_n$ are a sample from distribution F (CDF) if they are independent and all have distribution F.

Any function $g(X_1, X_2, \ldots, X_n)$ is a statistic defined on the sample. Such function is, on its own, a random variable.

Two most common statistics are:

Sample Mean: $$\overline{X} = \overline{X}_n = \frac{1}{n}\sum_{k=1}^{n} X_k$$

Sample Variance: $$S^2 = S_n^2 = \frac{1}{n-1}\sum_{k=1}^{n} (X_k - \overline{X}_n)^2$$

If F is a distribution with population mean μ and population variance σ^2, it can be shown that

$$E(\overline{X}_n) = \mu, \quad Var(\overline{X}_n) = \frac{\sigma^2}{n}, \quad \text{and} \quad E(S_n^2) = \sigma^2.$$

Sample Mean is an ***unbiased estimator*** of the *population mean*

Sample Variance is an ***unbiased estimator*** of the *population variance*

Sum of independent random variables

For any random variables X and Y, $E(X+Y) = E(X) + E(Y)$.

Moreover, for *independent* X and Y, $Var(X+Y) = Var(X) + Var(Y)$.

Slide 1.40: properties (5) and (6)

The identities extend to any finite sequence $X_1, X_2, \ldots, X_n$:

(E) $\quad E(X_1+X_2+\ldots+X_n) = E(X_1) + E(X_2) + \ldots + E(X_n).$

And for independent random variables:

(V) $\quad Var(X_1+X_2+\ldots+X_n) = Var(X_1) + Var(X_2) + \ldots + Var(X_n).$

Consequently, if $X_1, X_2, \ldots, X_n$ is a sample from the distribution F with mean μ and variance σ^2:

$$E(\overline{X}_n) = E\left(\frac{1}{n}\sum_{k=1}^{n} X_k\right) \overset{*}{=} \frac{1}{n} E\left(\sum_{k=1}^{n} X_k\right) = \frac{1}{n}\sum_{k=1}^{n} E(X_k) = \frac{1}{n}\cdot n\mu = \mu$$

(by (E) on the $= E\left(\sum_{k=1}^{n} X_k\right)$ step)

$$Var(\overline{X}_n) = Var\left(\frac{1}{n}\sum_{k=1}^{n} X_k\right) \overset{**}{=} \frac{1}{n^2} Var\left(\sum_{k=1}^{n} X_k\right) = \frac{1}{n^2}\sum_{k=1}^{n} Var(X_k) = \frac{1}{n^2}\cdot n\sigma^2 = \frac{\sigma^2}{n}$$

(by (V) on the $= Var\left(\sum_{k=1}^{n} X_k\right)$ step)

(*) property (2) on slide 1.40.

(**) property (4) on slide 1.40.

Examples

Example: Suppose that $X_1, X_2, \ldots, X_{100}$ is a sample from $N(0,25)$. What is the expected value and the variance of $\bar{X}_{100}$?

Solution: Population mean $\mu = 0$ and variance $\sigma^2 = 25$.

$$\text{Then } E(\bar{X}_{100}) = \mu = 0 \text{ and } Var(\bar{X}_{100}) = \frac{\sigma^2}{n} = \frac{25}{100} = 0.25$$

Example: Suppose that $X_1, X_2, \ldots, X_{144}$ is a sample from $Exp(0.25)$. What is the expected value and the variance of $3(\bar{X}_{144} - 4)$?

Solution: Population mean $\mu = \frac{1}{0.25} = 4$ and variance $\sigma^2 = \frac{1}{0.25^2} = 16$.

Thus $E(3(\bar{X}_{144} - 4)) = 3E(\bar{X}_{144} - 4) = 3(E(\bar{X}_{144}) - 4) = 3 \cdot 0 = 0$

and $Var(3(\bar{X}_{144} - 4)) = 3^2 Var(\bar{X}_{144} - 4) = 9\ Var(\bar{X}_{144}) = 9 \cdot \frac{16}{144} = 1$

Note: $3(\bar{X}_{144} - 4)$ is the ***Transformed Sample Mean*** of this sample:

$$\bar{Z}_n = \frac{\bar{X}_n - \mu}{\sqrt{\frac{\sigma^2}{n}}} = \frac{\bar{X}_{144} - 4}{\sqrt{\frac{16}{144}}} = \frac{\bar{X}_{144} - 4}{\frac{1}{3}} = 3(\bar{X}_{144} - 4)$$

Transformed Sample Mean

Note: If random variable X has a finite expected value μ and variance σ^2 then

$$Z = \frac{X - \mu}{\sigma}$$

is a random variable with expected value 0 and variance 1.

This follows easily from properties (1) – (4) on slide 1.40:

$$E(Z) = E\left(\frac{1}{\sigma}(X-\mu)\right) \overset{(2)}{=} \frac{1}{\sigma}E(X-\mu) \overset{(1)}{=} \frac{1}{\sigma}(\underbrace{EX}_{=\mu} - \mu) = 0, \text{ and}$$

$$Var(Z) = Var\left(\frac{1}{\sigma}(X-\mu)\right) \overset{(4)}{=} \frac{1}{\sigma^2}Var(X-\mu) \overset{(3)}{=} \frac{1}{\sigma^2}\underbrace{VarX}_{=\sigma^2} = 1.$$

If $X_1, X_2, \ldots, X_n$ is a sample from the distribution F with population mean μ and population variance σ^2, then we define the *Transformed Sample Mean*

$$Z_n = \frac{\bar{X}_n - \mu}{\frac{\sigma}{\sqrt{n}}} = \frac{\sqrt{n}}{\sigma}(\bar{X}_n - \mu)$$

Since $E(\bar{X}_n) = \mu$ and $Var(\bar{X}_n) = \frac{\sigma^2}{n}$ by the comment above Z_n has expected value 0 and variance 1.

Normal Samples

For independent random variables $X \sim N(\mu,\sigma^2)$ and $Y \sim N(\nu,\rho^2)$, $X+Y$ is a normal random variable ('*independent normal's are closed under summation*').

From (E) and (V) it follows that

$$X + Y \sim N(\mu+\nu,\ \sigma^2+\rho^2).$$

Similarly, for any finite sequence of independent normal's,

$$X_1+X_2+\ldots+X_n \sim N(\mu_1+\mu_2+\ldots+\mu_n,\ \sigma_1^2+\sigma_2^2+\ldots+\sigma_n^2).$$

With help of (2) and (4) on slide 1.40 we conclude that the *Sample Mean* of the sample $X_1, X_2, \ldots, X_n$ from $N(\mu, \sigma^2)$ is

$$\bar{X}_n \sim N\left(\mu, \frac{\sigma^2}{n}\right).$$

By *transformation to standard normal* (see (4), slide 1.41): $\dfrac{\bar{X}_n-\mu}{\frac{\sigma}{\sqrt{n}}} \sim N(0,1).$

Note: the latter is the *Transformed Sample Mean* and from the previous slide we know its expected value is zero and variance 1. However, for the **normal** sample more can be said: the Transformed Sample Mean is a standard **normal** variable.

Examples

Slide 1.42

Recall: Suppose $X \sim N(3, 2)$. What is the distribution of $2X - 5$?

Solution: $2X - 5 \sim N(1,8)$.

Example: Assume that $X \sim N(4,12)$ and $Y \sim N(2,8)$ are independent. What is the distribution of: (a) $X + Y$, (b) $2X - 3Y$.

Solution (a): Sum of independent normals is normal, so $X + Y$ is a normal random variable. Then $E(X + Y) = E(X) + E(Y) = 4 + 2 = 6$, and

$$Var(X + Y) = Var(X) + Var(Y) = 12 + 8 = 20.$$

Therefore $X + Y \sim N(6,20)$.

Solution (b): $2X$ and $-3Y$ are normal random variables (prop. (3) on 1.41) and they are also independent, so their sum is a normal random variable.

Then $E(2X - 3Y) = E(2X) + E(-3Y) = 2E(X) - 3E(Y) = 2 \cdot 4 - 3 \cdot 2 = 2$, and

$$Var(2X - 3Y) = Var(2X) + Var(-3Y) = 2^2 Var(X) + (-3)^2 Var(Y) = 4 \cdot 12 + 9 \cdot 8 = 120.$$

Therefore $2X - 3Y \sim N(2,120)$.

Examples

Example: Let $X_1, X_2, \ldots, X_{100}$ be a sample from $N(-10, 4)$.

(a) What is the distribution of its Sample Mean?

(b) What is the distribution of $5(\bar{X}_{100} + 10)$?

Solution (a): *Sample Mean* $\bar{X}_n$ of the sample from $N(\mu, \sigma^2)$ has $N\left(\mu, \frac{\sigma^2}{n}\right)$ distribution (see 2.11). In this case we have population mean $\mu = -10$, population variance $\sigma^2 = 4$, and sample size $n = 100$, therefore

$$\bar{X}_{100} \sim N\left(-10, \frac{4}{100}\right) = N\left(-10, \frac{1}{25}\right)$$

Solution (b): Notice that $5(\bar{X}_{100} + 10)$ is normal since $\bar{X}_{100}$ is normal.

$$E(5(\bar{X}_{100} + 10)) = 5E(\bar{X}_{100}) + 50 = 5 \cdot (-10) + 50 = 0, \text{ and}$$

$$Var(5(\bar{X}_{100} + 10)) = 5^2 Var(\bar{X}_{100} + 10) = 25\, Var(\bar{X}_{100}) = 25 \cdot \frac{1}{25} = 1.$$

Note: This is the Transformed Sample Mean of a sample $X_1, X_2, \ldots, X_{100}$:

$$5(\bar{X}_{100} + 10) = \frac{\bar{X}_{100} + 10}{\frac{1}{5}} = \frac{\bar{X}_{100} - (-10)}{\sqrt{\frac{1}{25}}} = \frac{\bar{X}_n - \mu}{\frac{\sigma}{\sqrt{n}}} \sim N(0,1)$$

How to "create" a sample in *R*?

We can use random number generators in *R*. But they produce sequences (vectors) of *numbers*, not random variables.

Idea: Generate a vector of N random numbers from a given distribution (use `rnorm()`, `runif()`, ...) and think of it as *a representative* of the first random variable, X_1, from the sample.

Repeat this for random variables X_2, … X_n.

$X_1^{(1)}$	$X_2^{(1)}$	$X_3^{(1)}$		$X_n^{(1)}$	
$X_1^{(2)}$	$X_2^{(2)}$	$X_3^{(2)}$		$X_n^{(2)}$	
$X_1^{(3)}$	$X_2^{(3)}$	$X_3^{(3)}$	…	$X_n^{(3)}$	
⋮	⋮	⋮		⋮	
$X_1^{(N)}$	$X_2^{(N)}$	$X_3^{(N)}$		$X_n^{(N)}$	
'representative' of	X_1	X_2	X_3	X_n	

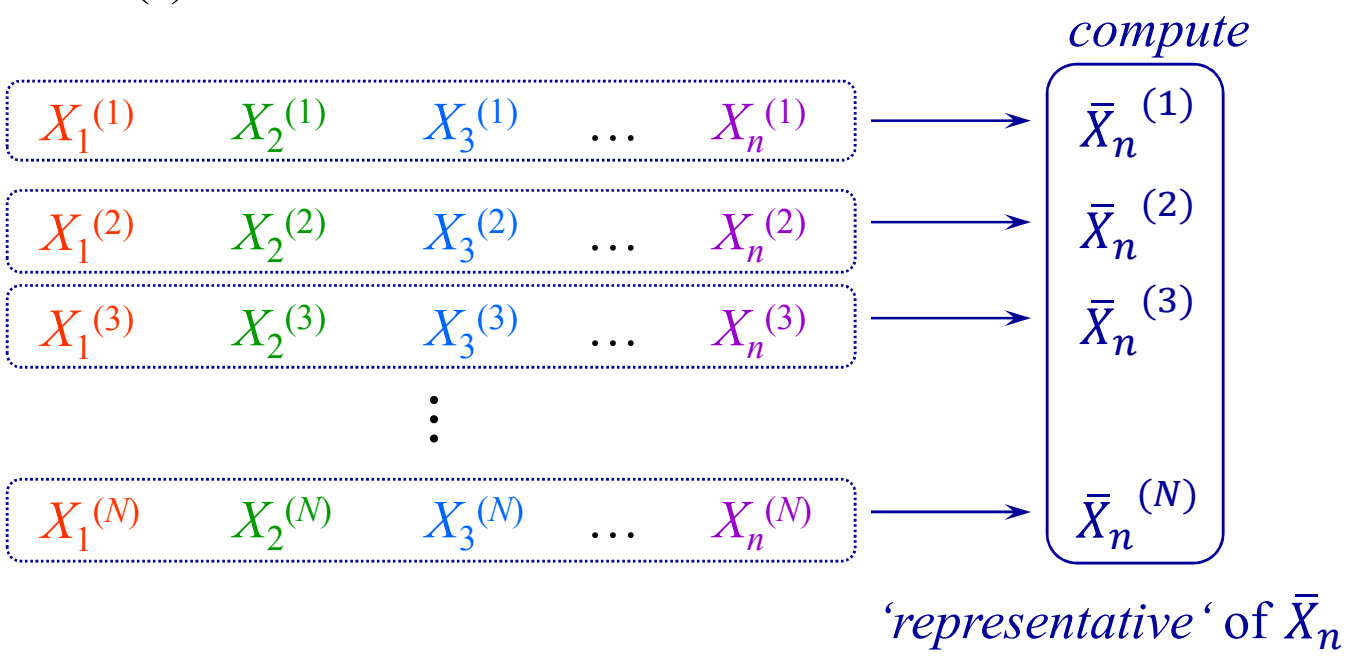
Length N of the 'representative' vector is typically going to be very large.

Sample size n is typically smaller.

Slight modification of the idea

With the sample representation as described, in order to compute the statistics of interest (sample mean, variance, median, …) we need to store $N \cdot n$ numbers in the memory. This is fine if n is a 'small' number.

If we put some faith in our random number generator, we can 'transpose' this procedure, i.e., generate the matrix row-by-row, and at each step calculate the desired statistic(s).



The illustration is for the *Sample Mean*, but can be mimicked for any desired statistic.

A brief detour into CLT

Cleary by taking larger N we obtain better representation of our sample, and consequently a better representation of the statistic(s) of interest.

But what happens as we increase the sample size n?

For the sample mean this is answered by *Central Limit Theorem*.

CLT: Let $X_1, X_2, \ldots$ be a sequence of i.i.d. random variables from the distribution F with population mean μ and variance σ^2. Then the sequence of random variables $Z_1, Z_2, \ldots$ defined by

Transformed Sample Means

$$Z_n = \frac{\bar{X}_n - \mu}{\frac{\sigma}{\sqrt{n}}} = \sqrt{n}\,\frac{\bar{X}_n - \mu}{\sigma} = \frac{1}{\sqrt{n}} \sum_{k=1}^{n} \frac{X_k - \mu}{\sigma}$$

converges *in distribution* to $N(0,1)$ as $n \to \infty$.

Convergence in distribution: if we denote by F_n the CDF of Z_n and the CDF of the standard normal by F then

$$\lim_{n \to \infty} F_n(x) = F(x), \text{ for all real numbers } x.$$

Sample Variance

Given a sample $X_1, X_2, \ldots, X_n$ from distribution F we define

$$\text{Sample Variance } S^2 = S_n^2 = \frac{1}{n-1}\sum_{k=1}^{n}(X_k - \bar{X}_n)^2$$

WHY? $n-1$

Recall: This is a random variable (like any statistic on sample).

Were the population mean μ known, then

$$E\left(\frac{1}{n}\sum_{k=1}^{n}(X_k-\mu)^2\right) = \frac{1}{n}E\left(\sum_{k=1}^{n}(X_k-\mu)^2\right) = \frac{1}{n}\sum_{k=1}^{n}E(X_k-\mu)^2$$

$$= \frac{1}{n}\sum_{k=1}^{n}Var(X_k) = \frac{1}{n}\,n\sigma^2 = \sigma^2.$$

But the population mean μ is unknown and our best estimate (*unbiased estimator*) for it is the *Sample Mean*.

$$E(\bar{X}_n) = \mu$$

Sample Variance cont.

Idea: we could try to estimate the population variance with

$$\frac{1}{n}\sum_{k=1}^{n}(X_k - \overline{X}_n)^2$$

Unfortunately $E\left(\frac{1}{n}\sum_{k=1}^{n}(X_k - \overline{X}_n)^2\right) = \frac{n-1}{n}\sigma^2.$ (simulations confirm this)

By multiplying both sides by $\frac{n}{n-1}$ we get

$$E(S_n^2) = E\left(\frac{1}{n-1}\sum_{k=1}^{n}(X_k - \overline{X}_n)^2\right) = \sigma^2.$$

Answer to WHY?: an unbiased estimator of population variance σ^2 is obtained when the sum of squares is divided by $n - 1$, not n.

Intuitive answer to WHY?

Given a sample $X_1, X_2, \ldots, X_n$ from distribution F with population mean μ, it is reasonable to expect that

$$\sum_{k=1}^{n}\left(X_k - \bar{X}_n\right)^2 \le \sum_{k=1}^{n}\left(X_k - \mu\right)^2.$$

In fact, given any n numbers $x_1, x_2, \ldots, x_n$ one can show that the function

$$f(u) = (x_1-u)^2 + (x_2-u)^2 + \ldots + (x_n-u)^2 =$$
$$= (u-x_1)^2 + (u-x_2)^2 + \ldots + (u-x_n)^2$$

attains minimum at point u that is the average of x_k's.

(basic calculus: take the derivative of $f(u)$, set it equal zero and solve for u)

Divide both sides by n, and notice that the inequality extends to expected values of those random variables above

$$E\left(\frac{1}{n}\sum_{k=1}^{n}\left(X_k - \bar{X}_n\right)^2\right) \le E\left(\frac{1}{n}\sum_{k=1}^{n}\left(X_k - \mu\right)^2\right) = \sigma^2.$$

The correction factor turns out to be $\frac{n}{n-1}$.

Degrees of Freedom

Given a sample $X_1, X_2, \ldots, X_n$ from distribution F the *residual* random variables

$$r_1 = X_1 - \bar{X}_n$$
$$r_2 = X_2 - \bar{X}_n$$
$$r_3 = X_3 - \bar{X}_n$$
$$\ldots$$
$$r_n = X_n - \bar{X}_n$$

are NOT independent. Why? Because they sum up to zero!

However, they are independent for any choice of $n - 1$ of them.

This yields a notion of $n - 1$ *degrees of freedom*!

Expected value of Sample Variance: *R*

Given a sample $X_1, X_2, \ldots, X_n$ from distribution F we define

$$\text{Sample Variance } S^2 = S_n^2 = \frac{1}{n-1}\underbrace{\sum_{k=1}^{n}(X_k - \bar{X}_n)^2}_{\Sigma} = \frac{\Sigma}{n-1}$$

Question: Why is Σ divided by $n - 1$ instead of n?

We can use *R* to estimate expected values of random variables $\frac{\Sigma}{n}$ and $\frac{\Sigma}{n-1}$:

	step 1	*step 2*	*step 3*	
$X_1^{(1)}\ X_2^{(1)}\ \ldots\ X_n^{(1)}$	$\bar{X}_n^{(1)}$	$\Sigma^{(1)} = \sum_{k=1}^{n}\left(X_k^{(1)} - \bar{X}_n^{(1)}\right)^2$	$\frac{\Sigma^{(1)}}{n}$	$\frac{\Sigma^{(1)}}{n-1}$
$X_1^{(2)}\ X_2^{(2)}\ \ldots\ X_n^{(2)}$	$\bar{X}_n^{(2)}$	$\Sigma^{(2)} = \sum_{k=1}^{n}\left(X_k^{(2)} - \bar{X}_n^{(2)}\right)^2$	$\frac{\Sigma^{(2)}}{n}$	$\frac{\Sigma^{(2)}}{n-1}$
⋮	⋮	⋮	⋮	⋮
$X_1^{(N)}\ X_2^{(N)}\ \ldots\ X_n^{(N)}$	$\bar{X}_n^{(N)}$	$\Sigma^{(N)} = \sum_{k=1}^{n}\left(X_k^{(N)} - \bar{X}_n^{(N)}\right)^2$	$\frac{\Sigma^{(N)}}{n}$	$\frac{\Sigma^{(N)}}{n-1}$

Our estimates for expected values of $\frac{\Sigma}{n}$ and $\frac{\Sigma}{n-1}$ are averages of these columns.

Confidence intervals?

Confidence interval is an open interval that with probability ***C*** includes an ***unknown*** population (numeric) property.

Probability ***C*** is called the confidence level.

Examples of a *population property:* population mean, population variance, population tenth percentile or third quartile, etc.

Classical Theory	Simulations in *R*
Confidence interval is calculated from the sample and almost certainly will be different for different samples. Relies on knowledge of the actual distribution for the statistics related to the population property of interest. A setback is that such distributions are easy to obtain only under very restrictive assumptions.	Create simulated samples and (from them) an empirical sampling distribution and use it to find empirical confidence interval. Few restrictions, yet accuracy depends on number of simulated samples and in each simulation slightly different distribution and interval will be obtained.

Probabilistic foundation

Confidence Interval for Normal sample Population Mean

If $X_1, X_2, \ldots, X_n$ is a sample from $N(\mu, \sigma^2)$ then:

Transformed Sample Mean

$$\frac{\overline{X}_n - \mu}{\frac{\sigma}{\sqrt{n}}} \sim N(0,1)$$

"known variance"

$$\frac{\overline{X}_n - \mu}{\frac{S_n}{\sqrt{n}}} \sim t(n-1)$$

"unknown variance"

Standard Error of a sample

The first claim we proved on lecture slide 2.11.

The second claim is much harder to prove.

Confidence intervals for population mean

Population mean confidence interval is an open interval which, with (probability) confidence level ***C***, includes an ***unknown*** population mean μ.

It is calculated from the sample and almost certainly will be different for different samples.

Recall: If $X_1, X_2, \ldots, X_n$ is a sample from $N(\mu, \sigma^2)$ and σ^2 is known then

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

Given a confidence level C (for example $C = 95\% = 0.95$),
find the point $z_C = z_{0.95}$ on the x-axis such that $P(-z_C < Z < z_C) = 0.95$.

If we can find such point we are done! Why?

Confidence intervals for population mean

$$-z_C < Z < z_C$$

$$\Leftrightarrow \; -z_C < \frac{\overline{X}_n - \mu}{\frac{\sigma}{\sqrt{n}}} < z_C$$

$$\Leftrightarrow \; -\frac{\sigma}{\sqrt{n}} z_C < \overline{X}_n - \mu < \frac{\sigma}{\sqrt{n}} z_C$$

$$\Leftrightarrow \; -\frac{\sigma}{\sqrt{n}} z_C - \overline{X}_n < -\mu < \frac{\sigma}{\sqrt{n}} z_C - \overline{X}_n$$

$$\Leftrightarrow \; \overline{X}_n + \frac{\sigma}{\sqrt{n}} z_C > \mu > \overline{X}_n - \frac{\sigma}{\sqrt{n}} z_C$$

How to interpret "with probability 0.95?"

Imagine you sample $X_1, X_2, \ldots, X_n$ from $N(\mu, \sigma^2)$ independently thousand times.

You get a thousand open intervals of same length centered at different points.

Population mean μ would be inside approximately 950 of those intervals, and outside of approximately 50 of them.

Hence $0.95 = P(-z_C < Z < z_C) = P\left(\overline{X}_n - \frac{\sigma}{\sqrt{n}} z_C < \mu < \overline{X}_n + \frac{\sigma}{\sqrt{n}} z_C\right)$

i.e., with probability 0.95 the population mean μ lies inside

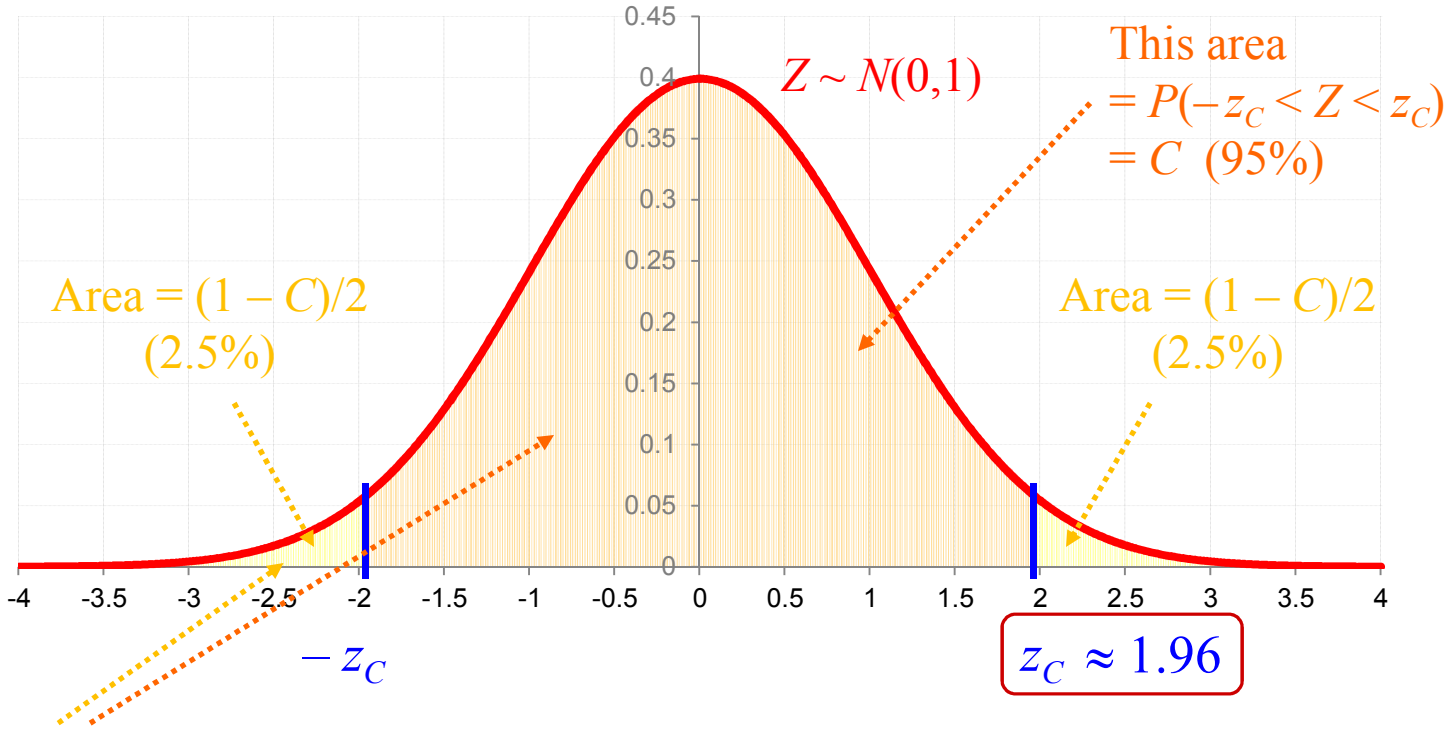
$$\left(\overline{X}_n - \frac{\sigma}{\sqrt{n}} z_C,\; \overline{X}_n + \frac{\sigma}{\sqrt{n}} z_C\right)$$

Confidence Interval for confidence C and given normal sample (known variance case)

Question: how do we compute z_C?

Given a confidence level C (e.g., $C = 95\% = 0.95$),
find the point z_C on the x-axis such that $P(-z_C < Z < z_C) = C$.

Idea:



These two areas $= P(Z < z_C) = F(z_C) = {}^{(1-C)}/_{2} + C = 2.5\% + 95\% = 97.5\%$

$$F(z_C) = 0.975 \iff z_C = F^{-1}(0.975) = \texttt{qnorm}(0.975) \approx 1.95996398454$$

Note: In order to compute z_C all we used was that $N(0,1)$ is a symmetric distribution.

Example

Numbers 7.85, 6.7, 9.83, 4, 8.61, 6.05, 5.39, 7.82, 5.63, 7.07, 11.2, and 6.8 are drawn from $N(\mu, 4)$. Calculate confidence interval for confidence level of 90%.

Solution: $n = 12$ and $\sigma^2 = 4$. Compute the sample mean:

$$\overline{X}_n = \frac{7.85 + 6.7 + \ldots + 6.8}{12} \approx 7.245833$$

For a conf. level $C = 90\% = 0.9$, $z_C = z_{0.9} =$ `qnorm(0.05+0.9)` ≈ 1.644854

$$\Rightarrow \frac{\sigma}{\sqrt{n}} \cdot z_C = \frac{2}{\sqrt{12}} \cdot z_{0.9} \approx 0.\ 9496567$$

Confidence interval: $\left(\overline{X}_n - \frac{\sigma}{\sqrt{n}} z_C, \overline{X}_n + \frac{\sigma}{\sqrt{n}} z_C\right)$

$$\approx (7.24583 - 0.9496567,\ 7.24583 + 0.9496567) = (6.296177,\ 8.195490)$$

Conclusion: with probability 0.9 the population mean μ is inside the interval (6.296177, 8.19549).

Confidence intervals: unknown σ

If $X_1, X_2, \ldots, X_n$ is a sample from $N(\mu, \sigma^2)$ and σ is NOT known then

$$T = \frac{\overline{X}_n - \mu}{\frac{S_n}{\sqrt{n}}} \sim t(n-1)$$

Student t distributions are symmetric!

Given a confidence level C (for example $C = 95\% = 0.95$),
find the point $t_C = t_{0.95}$ on the x-axis such that $P(-t_C < T < t_C) = 0.95$.

As before: $0.95 = P(-t_C < T < t_C) = P\left(\overline{X}_n - \frac{S_n}{\sqrt{n}} t_C < \mu < \overline{X}_n + \frac{S_n}{\sqrt{n}} t_C\right)$

i.e., with probability 0.95 the population mean μ lies inside

$$\left(\overline{X}_n - \frac{S_n}{\sqrt{n}} t_C, \overline{X}_n + \frac{S_n}{\sqrt{n}} t_C\right)$$

Confidence Interval for confidence C and given normal sample (unknown variance case)

Note: Unlike the known variance case, the lengths of the confidence intervals vary, since the *Sample Standard Deviation* S_n (unlike population standard deviation σ) depends on the sample.

Confidence intervals: unknown σ

Note: If you sample $X_1, X_2, \ldots, X_n$ from $N(\mu, \sigma^2)$ independently thousand times, you get thousand open intervals centered at different points but this time ***their lengths vary***, since the *Sample Standard Deviation* S_n depends on the sample.

Population mean μ will be inside approximately 950 of those intervals, and outside of approximately 50 of them.

To compute t_C we use the CDF of the *Student* $t(n-1)$ distribution.

For example, if the sample is of the size $n = 30$, then

$$F(t_C) = 0.975 \quad \Leftrightarrow \quad t_C = F^{-1}(0.975) = \texttt{qt}(0.975, 29) \approx 2.04522964213$$

Example

Numbers 7.85, 6.7, 9.83, 4, 8.61, 6.05, 5.39, 7.82, 5.63, 7.07, 11.2, and 6.8 are drawn from $N(\mu,\sigma^2)$. Calculate confidence interval for confidence level of 90%.

Solution: $n = 12$. Compute the sample mean and sample variance

$$\overline{X}_n = \frac{7.85 + 6.7 + \ldots + 6.8}{12} \approx 7.245833$$

$$S_n^2 = \frac{(7.85 - \overline{X}_n)^2 + (6.8 - \overline{X}_n)^2 + \ldots + (6.7 - \overline{X}_n)^2}{12 - 1} \approx 3.947008$$

For a conf. level $C = 90\% = 0.9$, $t_C = t_{0.9} =$ `qt(0.05+0.9,11)` ≈ 1.795885

$$\Rightarrow \frac{S_n}{\sqrt{n}} \cdot t_C \approx \frac{\sqrt{3.947008}}{\sqrt{12}} \cdot t_{0.9} \approx 1.029964$$

Confidence interval: $\left(\overline{X}_n - \frac{S_n}{\sqrt{n}} t_C, \overline{X}_n + \frac{S_n}{\sqrt{n}} t_C\right)$

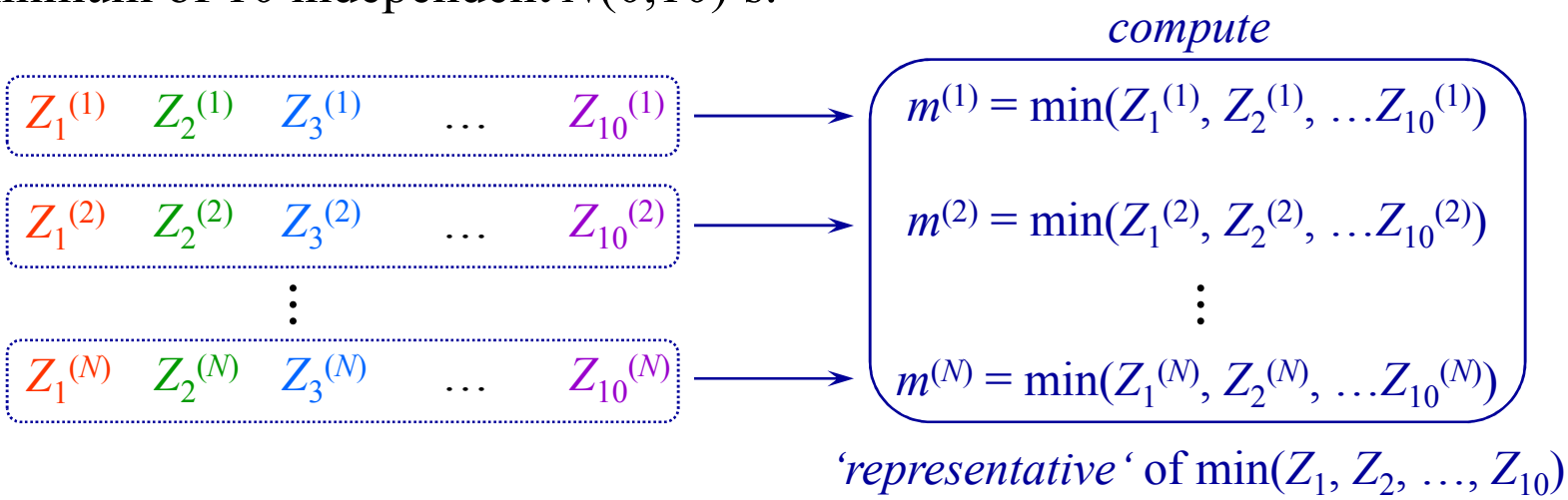
$\approx (7.24583 - 1.029964, 7.24583 + 1.029964) = (6.21587, 8.275797)$

Conclusion: with probability 0.9 the population mean μ is inside the interval (6.21587, 8.275797).

Empirical Confidence Interval example

Find the empirical 95% confidence interval for the minimum of the sample of size 10 from the Standard Normal distribution.

Idea: Use random number generator in *R* to generate a representative of a minimum of 10 independent $N(0,10)$'s:

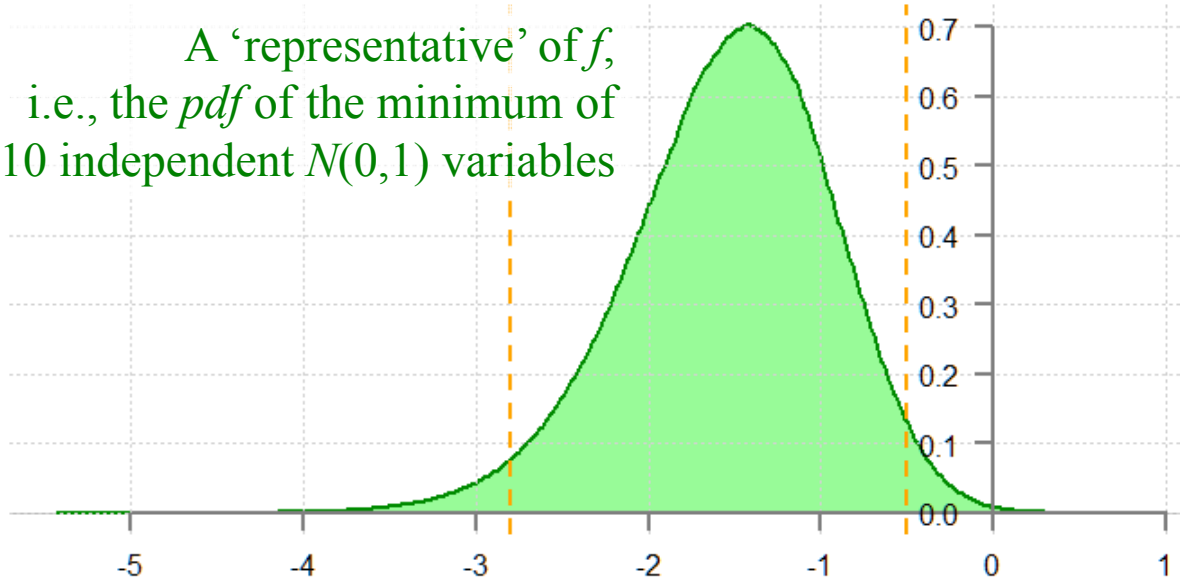


The obtained representative we call *sampling distribution vector*. For the "large" N's its empirical density will be an "accurate" representative of the true (theoretical) *pdf* of the minimum of ten independent $N(0,1)$'s.

Notation: Let f be the (*theoretical*) pdf of the minimum of ten independent $N(0,1)$ variables. Let F and Q be its CDF and Quantile functions, respectively.

Empirical Confidence Interval example (cont)

Empirical density of a sampling distribution vector of length one million:



Note: The solution to the problem is $CI = \big(Q(0.025), Q(0.975)\big)$, i.e., the 0.025 and 0.975 quantiles of *f*. The best we have is the sampling distribution vector whose *density()* is depicted above.

Consequently, the endpoints of the 95% confidence interval are obtained as 2.5^{th} and 97.5^{th} percentile of this vector (see in-class *R session*).

Hypothesis testing (intro)

A botanist is trying to determine whether an experimental growth substance applied to seeds affected the average height of plants. The standard average height is 306.9 mm. The botanist treated a random sample of 420 seeds with the extract and subsequently obtained the height data.

She decides to make a conclusion in the following way:

- If the standard mean of 306.9 is inside 95% confidence interval generated by her sample she will conclude that the treatment did not affect the growth.
- Otherwise, she will conclude the treatment did affect the growth.

The 95% confidence interval is (306.94, 307.53) and it does not include 306.9. Consequently the botanist concludes that the treatment affected plant growth.

Another take: Let μ denote the population mean of height of treated plant population.

A botanist makes a *hypothesis:* $\mu = 306.9$

and either accepts it, or rejects it – in favor of the opposite (*alternative*):

alternative hypothesis: $\mu \neq 306.9$

Note: $\mu_0 = 306.9$ is the *hypothesized value* for the pop. mean of the treated population. Notice that this value is likely a 'commonly known' fact among the botanists.

(Classical) Two-tailed t-test

Suppose we have a sample $X_1, X_2, \ldots, X_n$ from $N(\mu,\sigma^2)$ and the population parameters are not known. We make a hypothesis

H_0: $\mu = \mu_0$, where μ_0 is our guess[1].

This is called a null hypothesis. It could be an accepted fact based on past experience or it may represent a theory that has been put forward because it is believed to be true.

Its counterpart is the alternative hypothesis H_a: $\mu \neq \mu_0$

For example, in the botanist example the null hypothesis was that the substance did not affect the plant growth:

H_0: ***there is no difference*** between the mean height of treated and untreated plant populations,

H_a: ***there is a difference*** between the two.

A null hypothesis is either ***accepted*** or ***rejected***. How?

[1] as seen in botanist example, a guess could be a commonly known or accepted fact.

Two-tailed t-test (cont.)

Given a sample $X_1, X_2, \ldots, X_n$ from $N(\mu,\sigma^2)$ with unknown population parameters, compute the T statistic

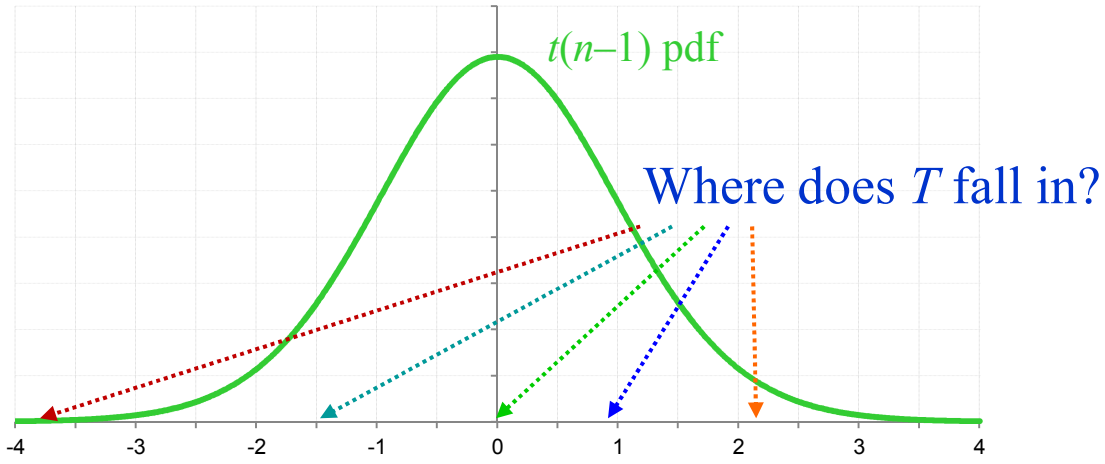
$$T = \frac{\bar{X}_n - \mu_0}{\frac{S_n}{\sqrt{n}}}.$$

Notice:

$$T = \frac{\bar{X}_n \boxed{-\mu+\mu} - \mu_0}{\frac{S_n}{\sqrt{n}}} = \underbrace{\boxed{\frac{\bar{X}_n - \mu}{\frac{S_n}{\sqrt{n}}}}}_{\text{has } t(n-1) \text{ distribution}} + \frac{\mu - \mu_0}{\frac{S_n}{\sqrt{n}}} \sim t(n-1) + \underbrace{\boxed{\frac{\sqrt{n}}{S_n}}}_{>0}(\mu - \mu_0)$$

If H_0: $\mu = \mu_0$ is true, then $T \sim t(n-1)$. Thus it is reasonable to establish where this numeric value resides relative to the pdf of $t(n-1)$ distribution:

Idea: If T came from $t(n-1)$ distribution, then it is more likely that it will be closer to zero.

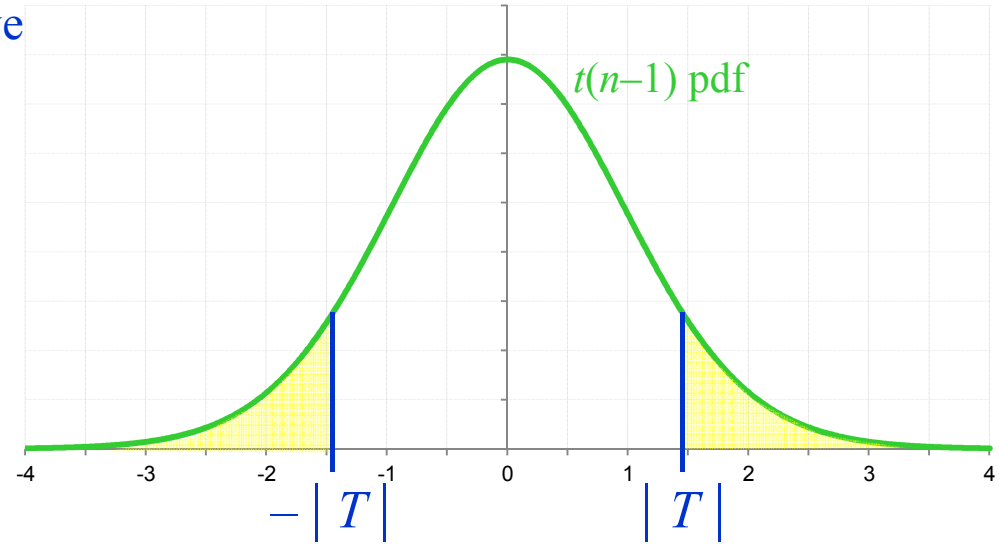


Two-tailed t-test (cont.)

Idea: 'measure' how likely is that the T value we computed is *that much away* from zero.

Note: Assume $\alpha = 0.05$; if we have 1000 samples from the normal distr. with population mean μ_0, for approx. 950 of those samples null hypothesis will be accepted.

However for approx. 50 ***it will be rejected***, despite the fact that the null hypothesis is correct!
This is often called ***Type 1 Error***.



That measure is exactly the sum of two yellow areas:

$$P\left(t(n-1) \le -|T|\right) + P\left(t(n-1) \ge |T|\right) = 2\,P\left(t(n-1) \le -|T|\right)$$

This is called the *p value*. If the p value is 'large', it is more likely that the T statistic value came from the $t(n-1)$ distribution, i.e., more likely that H_0 is true.

Thus if p value is larger than some agreed ***significance level α*** we ***accept the null hypothesis*** H_0. Typically $\alpha = 0.05$ (5%) or 0.01 (1%).

Summary: Two-tailed t-test

Given a sample $X_1, X_2, \ldots, X_n$ from $N(\mu,\sigma^2)$ with unknown population parameters, formulate the hypotheses

$$H_0: \mu = \mu_0 \qquad H_a: \mu \neq \mu_0$$

and compute:

T statistic $\quad T = \dfrac{\overline{X}_n - \mu_0}{\frac{S_n}{\sqrt{n}}}$

p value $\quad P\big(t(n-1) \leq -|T|\big) + P\big(t(n-1) \geq |T|\big)$

```
= pt(-abs(T),n-1) + 1 - pt(abs(T),n-1)
```

Given the agreed significance level α,

if p value $> \alpha$, ***accept the null hypothesis!***

if p value $\leq \alpha$, ***reject the null hypothesis in favor of the alternative!***

Example

Numbers 7.85, 6.7, 9.83, 4, 8.61, 6.05, 5.39, 7.82, 5.63, 7.07, 11.2, and 6.8 are drawn from $N(\mu,\sigma^2)$. Test the null hypothesis $\mu = 7$ for 10% significance level.

Solution: $\overline{X}_n = \dfrac{7.85 + 6.7 + \ldots + 6.8}{12} \approx 7.245833$

$$S_n^2 = \frac{(7.85 - \overline{X}_n)^2 + (6.8 - \overline{X}_n)^2 + \ldots + (6.7 - \overline{X}_n)^2}{12 - 1} \approx 3.947008$$

$$T = \frac{\overline{X}_n - \mu_0}{\frac{S_n}{\sqrt{n}}} \approx \frac{7.245833 - 7}{\sqrt{\frac{3.947008}{12}}} \approx 0.4286446$$

p value = `pt(-T,n-1) + 1-pt(T,n-1)`

≈ `pt(-0.4286446,11) + 1-pt(0.4286446,11)`

$\approx 0.6764553 > 0.1$

Conclusion: At 10% significance accept the null hypothesis $\mu = 7$.

Two-tailed t-test and Confidence Interval

Given a sample $X_1, X_2, \ldots, X_n$ from $N(\mu,\sigma^2)$ with unknown population parameters, then

$$\frac{\overline{X}_n - \mu}{\frac{S_n}{\sqrt{n}}} \sim t(n-1).$$

For confidence $C = 0.95$ find the point $t_{0.95}$ on the x-axis such that

$$P(-t_{0.95} < t(n-1) < t_{0.95}) = P(|t(n-1)| < t_{0.95}) = 0.95.$$

Then the Confidence Interval for the population mean μ is

$$\left(\overline{X}_n - \frac{S_n}{\sqrt{n}} t_{0.95},\ \overline{X}_n + \frac{S_n}{\sqrt{n}} t_{0.95}\right), \quad \text{i.e.,}$$

$$P\left(\overline{X}_n - \frac{S_n}{\sqrt{n}} t_{0.95} < \mu < \overline{X}_n + \frac{S_n}{\sqrt{n}} t_{0.95}\right) = 0.95.$$

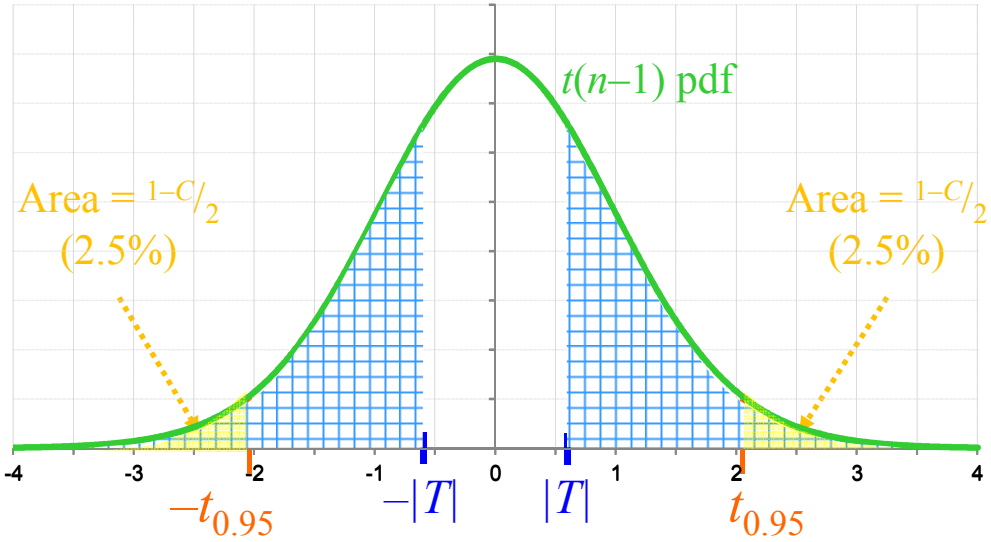
Take μ_0 from this interval and formulate the null hypothesis H_0: $\mu = \mu_0$, then

$$\overline{X}_n - \frac{S_n}{\sqrt{n}} t_{0.95} < \mu_0 < \overline{X}_n + \frac{S_n}{\sqrt{n}} t_{0.95} \quad \Leftrightarrow -t_{0.95} < \boxed{\frac{\overline{X}_n - \mu_0}{\frac{S_n}{\sqrt{n}}}} < t_{0.95}$$

'our' T statistic

Two-tailed t-test and Confidence Interval

Recap: If μ_0 is from the confidence interval then $-t_{0.95} < T < t_{0.95}$.



But then the p value is certainly larger than $0.05 = 1 - 0.95$.

Thus if we take any μ_0 from the confidence interval generated by confidence level C, the null hypothesis will be accepted at significance level $\alpha = 1 - C$.

One can easily show the converse: if the null hypothesis H_0: $\mu = \mu_0$ is accepted at significance level α then μ_0 will be in the confidence interval generated by the confidence level $C = 1 - \alpha$.

Alternative formulation for two-tailed t-test

Given a sample $X_1, X_2, \ldots, X_n$ from $N(\mu,\sigma^2)$ with unknown population parameters, formulate the hypotheses

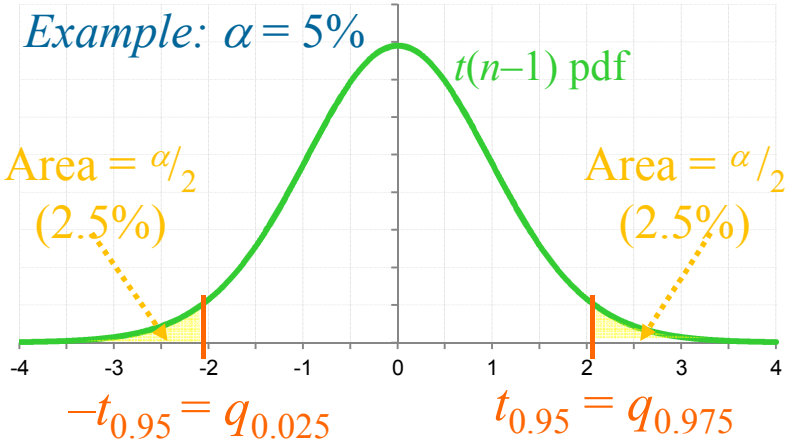
$$H_0: \mu = \mu_0 \qquad H_a: \mu \neq \mu_0$$

and compute

$$T \text{ statistic} \quad T = \frac{\bar{X}_n - \mu_0}{\frac{S_n}{\sqrt{n}}}$$

For significance level α compute

$$t_{1-\alpha} = \texttt{qt}(1 - \alpha/2, n - 1)$$



Then:

If $|T| < t_{1-\alpha}$, ***accept*** H_0 $\qquad (-t_{1-\alpha} < T < t_{1-\alpha})$

If $|T| \geq t_{1-\alpha}$, ***reject*** H_0 ***in favor of*** H_a $\qquad (T \leq -t_{1-\alpha}$ or $T \geq t_{1-\alpha})$

Note: This formulation ignores the *p value* altogether.

Upper-tailed t-test (motivation)

Salinity is a measure of the concentration of dissolved salts in water, commonly defined in "grams of salt in 1 kilogram of water." An environmentalist is collecting samples on the lake near the ocean. A recent hurricane caused the high tide and it is necessary to determine whether the salt content in the lake has increased significantly (with significance of 5% in the usual statistical testing sense). In the process 58 samples are taken and their salinity is measured. Use the *t-test* to determine whether the water salinity has increased significantly if the normal salinity of the lake is 0.5.
Note: the salinity of the ocean water is typically around 35.

Notice that the statement of the problem indicates that the salinity, had it changed, **should have increased**. The goal of the environmentalist is the same:

to show whether the salinity **truly increased** or **not**!

The two-tailed test with $H_0: \mu = 0.5$ and $H_a: \mu \neq 0.5$ can only try to answer whether the salinity changed or not.

If we try to set the test with $H_0: \mu \geq 0.5$ and $H_a: \mu < 0.5$, if the null-hypothesis is accepted it still includes the possibility of salinity staying the same ('=' is a part of '≥')

The 'test' we need is: $H_0: \mu \leq 0.5$ and $H_a: \mu > 0.5$.

The environmentalist will be convinced if the null-hypothesis is rejected.

Upper-tailed t-test (math)

Given a sample $X_1, X_2, \ldots, X_n$ from $N(\mu,\sigma^2)$ with unknown population parameters, recall that

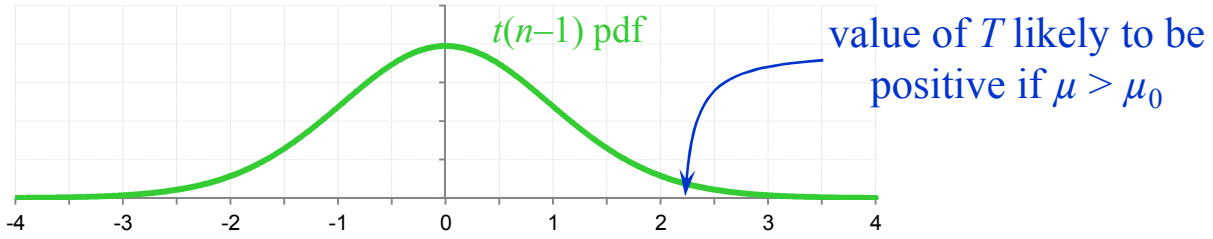
has $t(n-1)$ distribution

> 0

$$T = \frac{\overline{X}_n - \mu_0}{\frac{S_n}{\sqrt{n}}} = \frac{\overline{X}_n - \mu + \mu - \mu_0}{\frac{S_n}{\sqrt{n}}} = \boxed{\frac{\overline{X}_n - \mu}{\frac{S_n}{\sqrt{n}}}} + \frac{\mu - \mu_0}{\frac{S_n}{\sqrt{n}}} \sim t(n-1) + \boxed{\frac{\sqrt{n}}{S_n}}(\mu - \mu_0)$$

for any 'estimate' μ_0 of the population mean μ.

Cleary if $\mu = \mu_0$ the T statistic value will have $t(n-1)$ distribution. But if $\mu > \mu_0$ its value will (tend to) be positive, since the second term is positive.



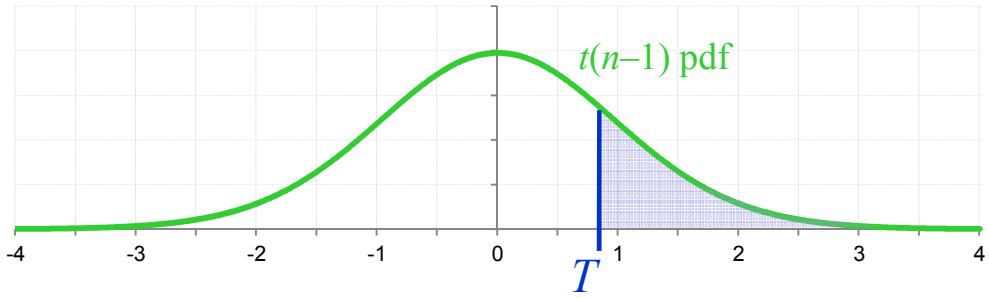
Accordingly, we construct the test with the following hypotheses:

$$H_0: \mu \le \mu_0 \qquad H_a: \mu > \mu_0$$

Note: If you intend to show $\mu > \mu_0$ you construct a test whose ***alternative hypothesis*** is exactly that!

Upper-tailed t-test

Back to the salinity example: the environmentalist is only interested in finding that the (new) population mean is truly larger than $\mu_0 = 0.5$ g/l. From the data the *T statistic* value is computed and this value is placed 'under' the $t(n-1)$ pdf.



If the T value happens to be negative or positive but 'not too large', there is no justification to the belief that μ is truly larger than $\mu_0 = 0.5$.

In other words, we should be accepting the null hypothesis H_0: $\mu \leq \mu_0$.

But how do we quantify 'too large'? We can measure the area to the right of T ('upper tail' of the distribution). That area is the upper-tailed test *p value*.

$$p \text{ value: } P\big(t(n-1) \geq T\big) = 1 - pt(T, n-1)$$

Given the agreed significance level α,

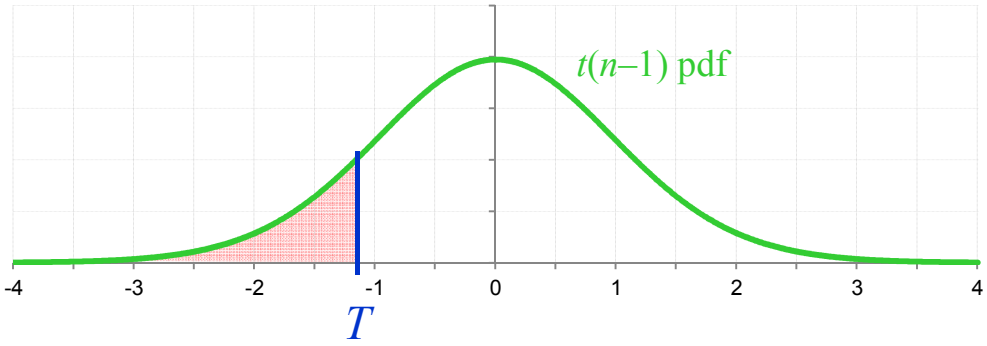
if p value $> \alpha$, accept H_0: $\mu \leq \mu_0$ (effectively $\mu = \mu_0$)

if p value $\leq \alpha$, reject H_0 in favor of H_a: $\mu > \mu_0$

Lower-tailed t-test

Following the same idea: we are trying to establish that the population mean is smaller than some value. E.g., "the average height … is less than 71 inches."

Accordingly we construct the test: H_0: $\mu \geq \mu_0$ $\quad H_a$: $\mu < \mu_0$



If the T value happens to be positive or negative but 'not too large', we have no reason to doubt that the population mean μ is truly not smaller than μ_0.

We quantify 'too large negative' as the area to the left of T ('lower tail' of the distribution), which is the lower-tailed test *p value:*

$$p \text{ value: } P\big(t(n-1) \leq T\big) = pt(T, n\text{–}1)$$

Given the agreed significance level α,

if p value $> \alpha$, accept H_0: $\mu \geq \mu_0$ (effectively $\mu = \mu_0$)

if p value $\leq \alpha$, reject H_0 in favor of H_a: $\mu < \mu_0$

Lower-tailed t-test alternative formulation

Given a sample $X_1, X_2, \ldots, X_n$ from $N(\mu,\sigma^2)$ with unknown population parameters, formulate the hypotheses

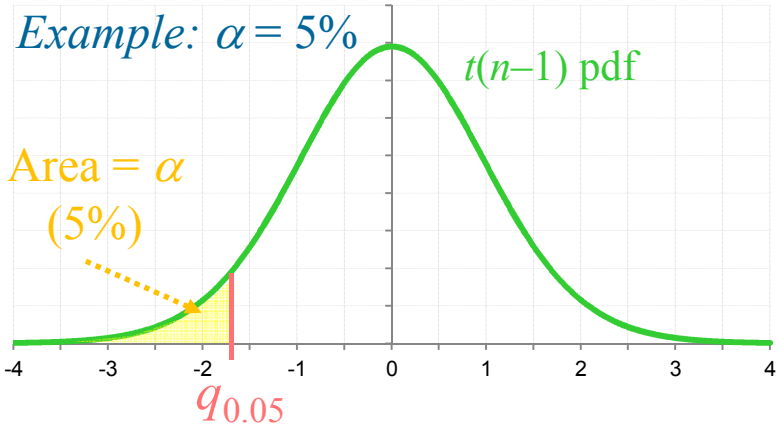
$$H_0: \mu \geq \mu_0 \qquad H_a: \mu < \mu_0$$

and compute T statistic

$$T = \frac{\overline{X}_n - \mu_0}{\frac{S_n}{\sqrt{n}}}$$

For significance level α compute

$$q_\alpha = \texttt{qt}(\alpha, n-1)$$



Then:

If $T > q_\alpha$, accept $H_0: \mu \geq \mu_0$ (effectively $\mu = \mu_0$)

If $T \leq q_\alpha$, reject H_0 in favor of $H_a: \mu < \mu_0$

Note: This formulation ignores the *p value*.

Upper-tailed t-test alternative formulation

Given a sample $X_1, X_2, \ldots, X_n$ from $N(\mu,\sigma^2)$ with unknown population parameters, formulate the hypotheses

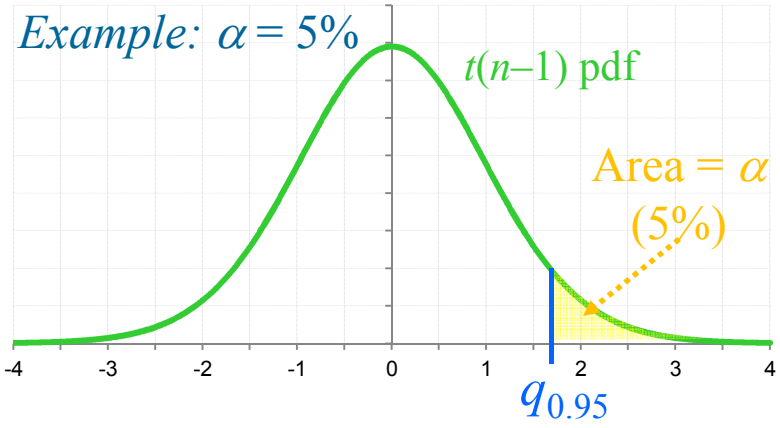
$$H_0: \mu \le \mu_0 \qquad H_a: \mu > \mu_0$$

and compute T statistic

$$T = \frac{\overline{X}_n - \mu_0}{\frac{S_n}{\sqrt{n}}}$$

For significance level α compute

$$q_{1-\alpha} = \texttt{qt}(1-\alpha, n-1)$$



Then:

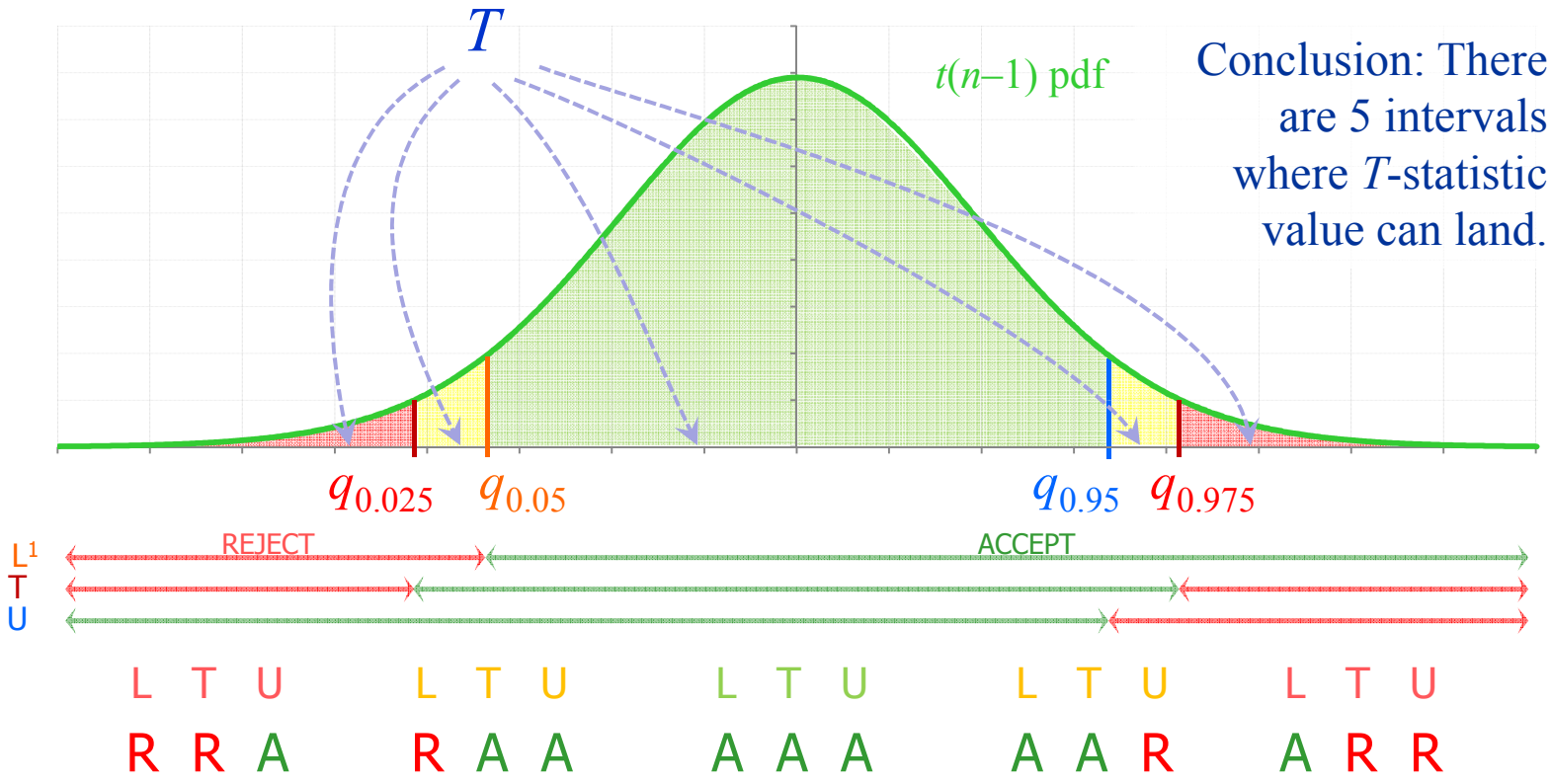
If $T < q_{1-\alpha}$, accept $H_0: \mu \le \mu_0$ (effectively $\mu = \mu_0$)

If $T \ge q_{1-\alpha}$, reject H_0 in favor of $H_a: \mu > \mu_0$

Note: This formulation ignores the *p value*.

Three tests combined

Given a sample $X_1, X_2, \ldots, X_n$ from $N(\mu,\sigma^2)$ with unknown population parameters, what are the possible outcomes of all three t-tests if we run them simultaneously? Namely, we compute T statistic based on assumed μ_0 and the sample and use the same significance level α (for simplicity assume $\alpha = 0.05$).



[1] L: Lower-tailed test T: Two-tailed test U: Upper-tailed test

Example

Salinity is a measure of the concentration of dissolved salts in water, commonly defined in "grams of salt in 1 kilogram of water." An environmentalist is collecting samples on the lake near the ocean. A recent hurricane caused the high tide and it is necessary to determine whether the salt content in the lake has increased significantly (with significance of 5% in the usual statistical testing sense). In the process 58 samples are taken and their salinity is measured. Use the *t-test* to determine whether the water salinity has increased significantly if the normal salinity of the lake is 0.5.

Notice that the statement of the problem indicates that the salinity, had it changed, should have increased. The goal of the environmentalist is the same: to show whether the salinity increased or not!

Thus we apply the upper-tailed test with $H_0: \mu \leq 0.5$ and $H_a: \mu > 0.5$.

R output:

```
t.test(u$Salinity, mu=0.5, alternative="greater")

        One Sample t-test

data:  u$Salinity
t = 1.7207, df = 57, p-value = 0.04537
alternative hypothesis: true mean is greater than 0.5
95 percent confidence interval:  0.5007201  Inf
sample estimates:
mean of x:  0.5254655
```

Conclusion: Reject the null-hypothesis in favor of the alternative: $\mu > 0.5$

Answer: the salinity **has increased**!

Question: what would two-tailed test yield?

The two-tailed test, however, accepts its null-hypothesis:

$$H_0: \mu = 0.5 \qquad H_a: \mu \neq 0.5$$

R output:

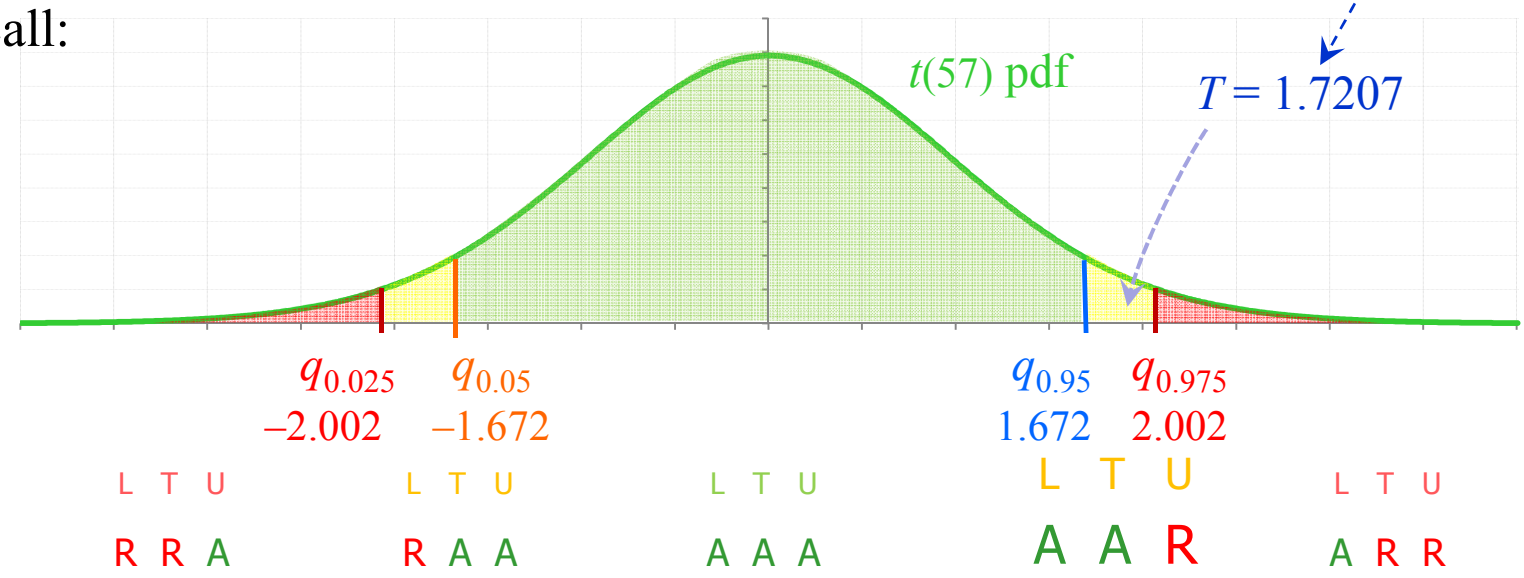
```
t.test(u$Salinity, mu=0.5)

t = 1.7207, df = 57, p-value = 0.09073
alternative hypothesis: true mean is not equal to 0.5
95 percent confidence interval:  0.4958298  0.5551013
```

Conclusion: Accept the null-hypothesis at 5% significance: $\mu = 0.5$

Answer: the salinity **has not changed**!

Recall:



The botanist example revisited

A botanist is trying to determine whether an experimental growth substance applied to seeds affected the average height of plants. The standard average height is 306.9 mm. The botanist treated a random sample of 420 seeds with the extract and subsequently obtained the height data.
Lets apply three t-tests at 5% significance level:

The botanist is secretly hoping to show that the substance increases plants' heights, thus we first apply the upper-tailed test with

$$H_0: \mu \leq 306.9 \quad \text{and} \quad H_a: \mu > 306.9.$$

R output:

```
t.test(ph, mu=306.9, alternative="greater")

        One Sample t-test

data:  ph
t = 2.2557, df = 419, p-value = 0.0123
alternative hypothesis: true mean is greater than 306.9
95 percent confidence interval:  306.9911  Inf
sample estimates:
mean of x:  307.2383
```

Conclusion: At 5% significance the botanist rejects the null-hypothesis in favor of the alternative: $\mu > 306.9$

The substance helps the plants grow!

Check

Dig in a bit more: The two-tailed test rejects its null-hypothesis as well:

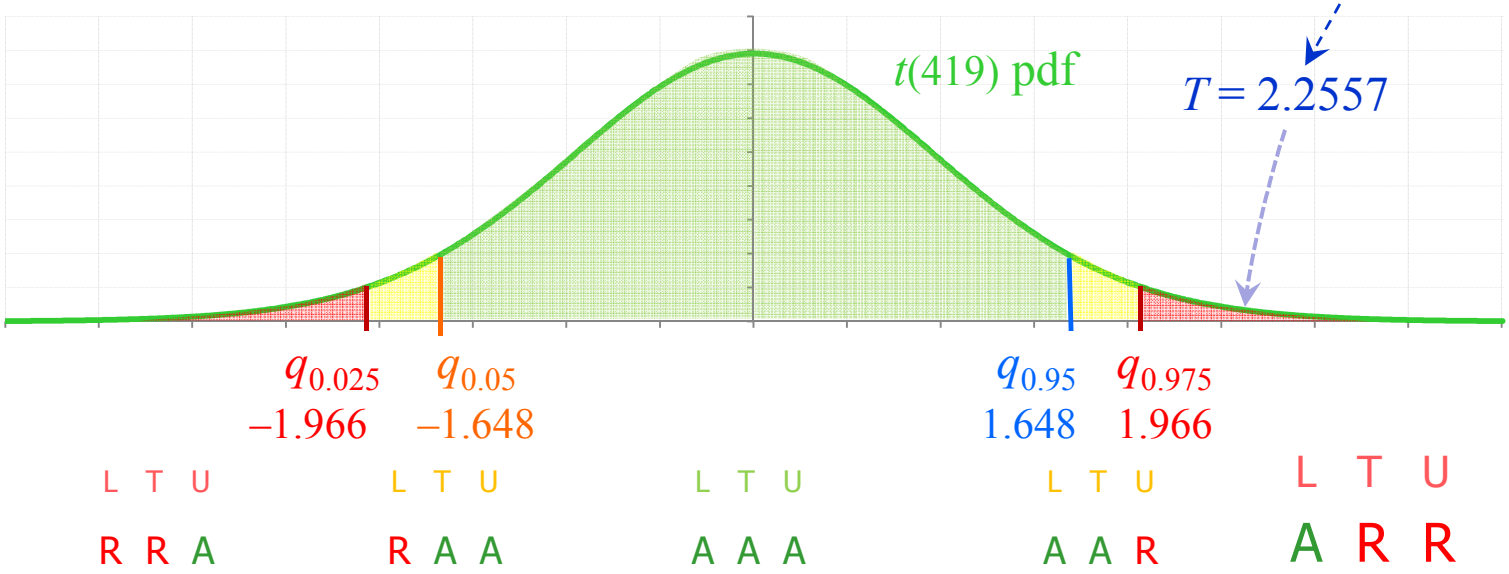
$$H_0: \mu = 306.9 \qquad H_a: \mu \neq 306.9$$

R output:

```
t.test(ph, mu=306.9)
…
t = 2.2557, df = 419, p-value = 0.0246
alternative hypothesis: true mean is not equal to 306.9
95 percent confidence interval:  306.9435  307.5332
```

Conclusion: Reject the null-hypothesis at 5% significance.

The substance has effect on the plant growth!



Rule of thumb

Two-tailed test should be used whenever our intention is to establish (within given significance level) that the population mean is equal to some hypothesized value. The null-hypothesis is either accepted or rejected in favor the alternative: population mean is not equal to the hypothesized value.

Using two-tailed test is the cautious way to conduct the hypothesis testing.

If an additional information (physics, past experience, sound intuition, etc.) indicates that the population mean is either strictly greater or strictly less than the hypothesized value, one of the one-tailed tests can be used:

(U) If the additional information indicates that the population mean is greater than the hypothesized value, use the upper-tailed test. The null-hypothesis is either accepted (in which case we effectively conclude that population mean is equal to the hypothesized value) or rejected in favor the alternative:

population mean is **strictly greater than** the hypothesized value.

(L) If the additional information indicates that the population mean is smaller than the hypothesized value, use the lower-tailed test. The null-hypothesis is either accepted (in which case we effectively conclude that population mean is equal to the hypothesized value) or rejected in favor the alternative:

population mean is **strictly less than** the hypothesized value.

Two-tailed z-test

Given a sample $X_1, X_2, \ldots, X_n$ from $N(\mu,\sigma^2)$ with unknown population mean and a known population variance σ^2, formulate the hypotheses

$$H_0: \mu = \mu_0 \qquad H_a: \mu \neq \mu_0$$

and compute:

Z statistic $\quad Z = \dfrac{\overline{X}_n - \mu_0}{\frac{\sigma}{\sqrt{n}}}$ Recall: If H_0 is true, then $Z \sim N(0,1)$

p value $\quad P\big(N(0,1) \leq -|Z|\big) + P\big(N(0,1) \geq |Z|\big)$

```
= pnorm(- abs(Z)) + 1 - pnorm(abs(Z))
```

Given the agreed significance level α,

if p value $> \alpha$, accept H_0

if p value $\leq \alpha$, reject H_0 in favor of H_a

Derive on your own the one-tailed versions of the *z*-test, as well as the alternative formulations analogous to those given for the *t*-tests.

One-sided Confidence interval

Population mean upper-tailed confidence interval is an open interval $(a, +\infty)$ which with confidence level ***C*** includes an ***unknown*** population mean μ.

Recall: If $X_1, X_2, \ldots, X_n$ is a sample from $N(\mu, \sigma^2)$ and σ^2 is known then

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

Given a confidence level C (e.g., $C = 95\%$), we have $P(Z < q_{0.95}) = 0.95$ for

$q_{0.95} =$ `qnorm(0.95)` (95^{th} percentile of $N(0,1)$)

Similarly as on slide 2.25 we can show: $Z < q_C$ *if and only if* $\mu > \overline{X}_n - \frac{\sigma}{\sqrt{n}} q_C$

i.e., $0.95 = P(Z < q_C) = P\left(\overline{X}_n - \frac{\sigma}{\sqrt{n}} q_C < \mu\right)$

Upper-tailed Confidence Interval for conf. C (known variance case)

i.e., with prob. 0.95 the population mean μ lies inside $\left(\overline{X}_n - \frac{\sigma}{\sqrt{n}} q_C, +\infty\right)$

Similarly a lower-tailed confidence interval can be derived, as well as both upper and lower tailed intervals for the unknown variance case.

Two-sample t-tests

Purpose: comparison of population means of two samples.

Two distinct cases: *paired samples* and *independent samples*.

Examples: (1) Randomly select 50 women and 50 men and record the average time (per week) they spend watching TV.

(2) Randomly select 50 couples and record the average time (per week) the wives spend watching TV, and then the same data for the husbands.

In a paired sample each entity is measured twice, resulting in pairs of observations. Common applications include *case-control studies:* a famous one was the demonstration of the link between tobacco smoking and lung cancer.

The tests are fundamentally different for these two cases. For example, for paired samples clearly the sample sizes must be equal. For independent samples this is not the requirement.

Paired samples tests

Paired samples: The sample selected from the first population is *related* to the corresponding sample from the second population.

Idea: Given two samples, $X_1, X_2, \ldots, X_n$ and $Y_1, Y_2, \ldots, Y_n$, consider their differences

$$D_1 = X_1 - Y_1, \quad D_2 = X_2 - Y_2, \ldots, \quad D_n = X_n - Y_n.$$

Then $D_1, D_2, \ldots, D_n$ is a sample (they are independent r.v.'s) from some distribution with population mean μ and variance σ^2. If that happens to be $N(\mu,\sigma^2)$, or if the sample size is large enough, we can use any classical (one-sample) test; if H_0: $\mu_1 = \mu_2$ is true, then

$$Z = \frac{\overline{D}_n}{\frac{\sigma}{\sqrt{n}}} \sim N(0,1) \quad \text{and} \quad T = \frac{\overline{D}_n}{\frac{S_n}{\sqrt{n}}} \sim t(n-1)$$

(S_n is the sample standard deviation of the 'difference sample')

Note: The only assumption we need is that the differences are normally distributed, or that the samples are large enough (CLT).

Independent samples tests

Independent samples: The sample from the first population is *independent* of the sample from the second population.

Idea: if we have two samples, $X_1, X_2, \ldots, X_n$ from $N(\mu_1, \sigma_1^2)$ and $Y_1, Y_2, \ldots, Y_m$ from $N(\mu_2, \sigma_2^2)$, then

$$\overline{X}_n \sim N\left(\mu_1, \frac{\sigma_1^2}{n}\right) \quad \text{and} \quad \overline{Y}_m \sim N\left(\mu_2, \frac{\sigma_2^2}{m}\right)$$

Note: By the CLT, the normality of the samples is not required if both samples are large enough.

and furthermore they are independent, thus their difference

$$\overline{X}_n - \overline{Y}_m \sim N\left(\mu_1 - \mu_2, \frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}\right).$$

Consequently if the population variances are known, the *z*-test with the null hypothesis H_0: $\mu_1 = \mu_2$ and the Z statistic

$$Z = \frac{\overline{X}_n - \overline{Y}_m}{\sqrt{\frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}}}$$ is straightforward.

(and both one-tailed *z*-tests as well)

Independent samples t-test (cont)

Problem: Unknown population variances.

Recall: Classical *t*-test was based on the theoretical result: for a normal sample (or, by CLT, for sample large enough),

$$\frac{\overline{X}_n - \mu}{\frac{S_n}{\sqrt{n}}} \sim t(n-1).$$

Similar result for differences of two independent sample means?

Somewhat: there are two subcases to consider:

Pooled Variances: Sample sizes are "nearly equal" and neither sample standard deviation is more than twice the other, i.e.,

$$\frac{1}{2} \leq \frac{S_1}{S_2} \leq 2.$$

Separate Variances: All other cases (*Welch test*).

Independent samples t-test (cont)

Pooled Variances: If the null hypothesis H_0: $\mu_1 = \mu_2$ is true, then

$$\frac{\bar{X}_n - \bar{Y}_m}{S_p\sqrt{\frac{1}{n}+\frac{1}{m}}} \sim t(n+m-2) \quad \text{with} \quad S_p = \sqrt{\frac{(n-1)S_n^2 + (m-1)S_m^2}{n+m-2}}.$$

Separate Variances: If the null hypothesis H_0: $\mu_1 = \mu_2$ is true, then

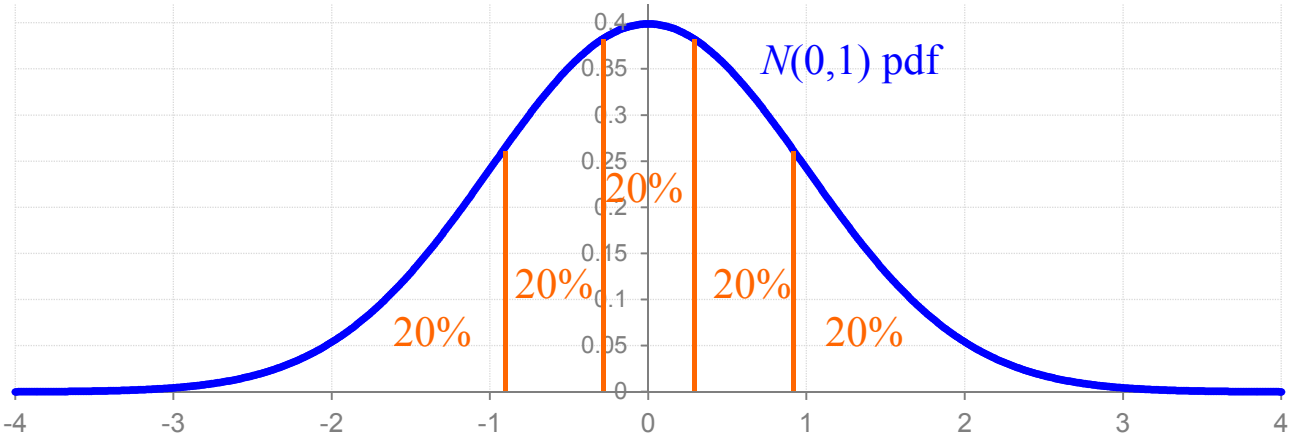
$$\frac{\bar{X}_n - \bar{Y}_m}{\sqrt{\frac{S_n^2}{n}+\frac{S_m^2}{m}}} \sim t(df) \quad \text{with } df = \frac{(n-1)\cdot(m-1)}{(m-1)C^2 + (1-C)^2(n-1)} \quad \& \quad C = \frac{\frac{S_n^2}{n}}{\frac{S_n^2}{n}+\frac{S_m^2}{m}}.$$

An alternate (conservative) option is to use $df = \min(n,m) - 1$.

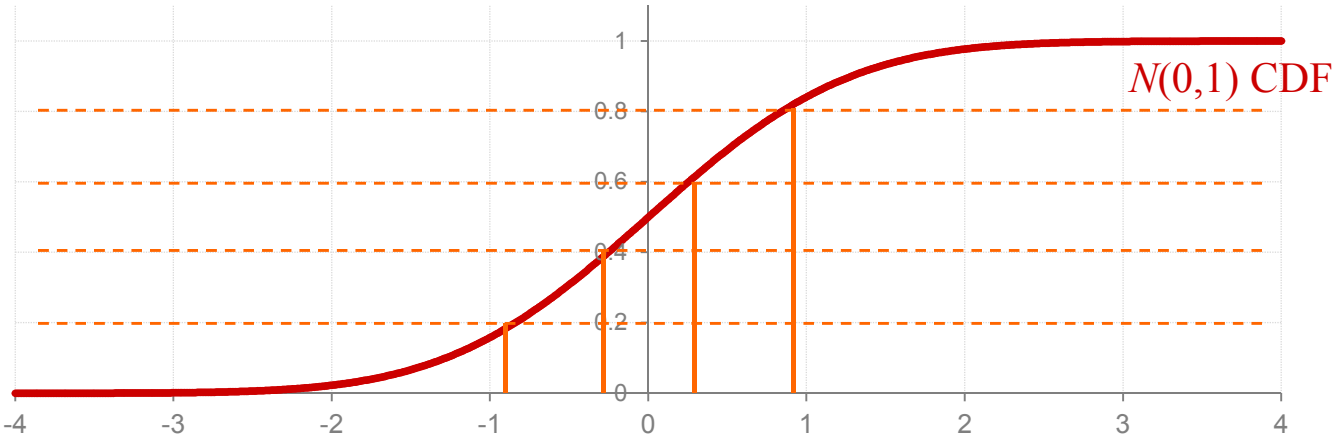
We will rely on *R* built in *t-test* function for applications of these tests.

Q: Is sample from Normal distribution?

Idea: Generate five random numbers from *Standard Normal* distribution.

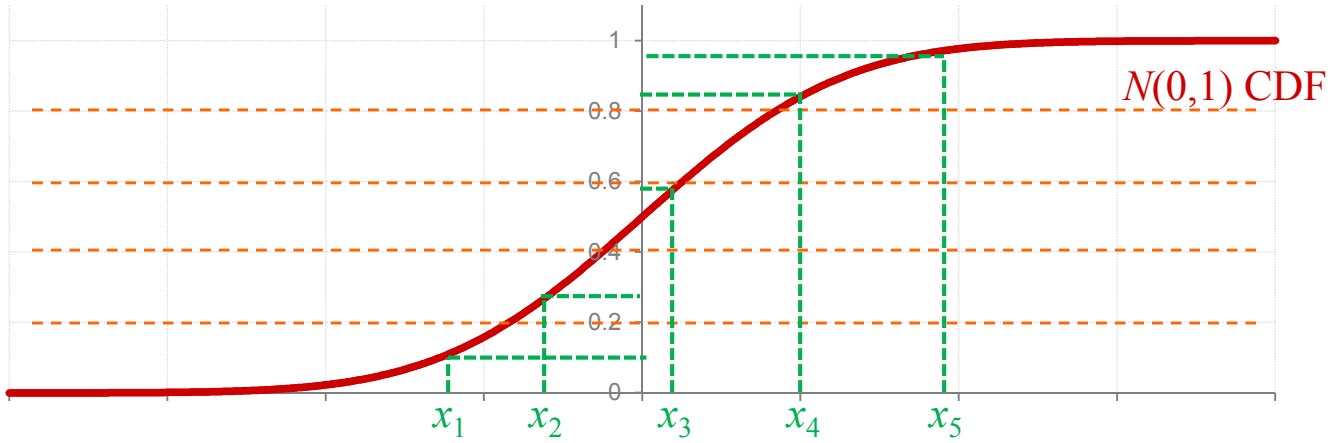


Where should they 'fall in'? If we were to repeat generation many times, they should "uniformly" fill the five intervals above!



Is sample from Normal distribution? (cont)

We can sort the five random numbers and compute their CDF values.



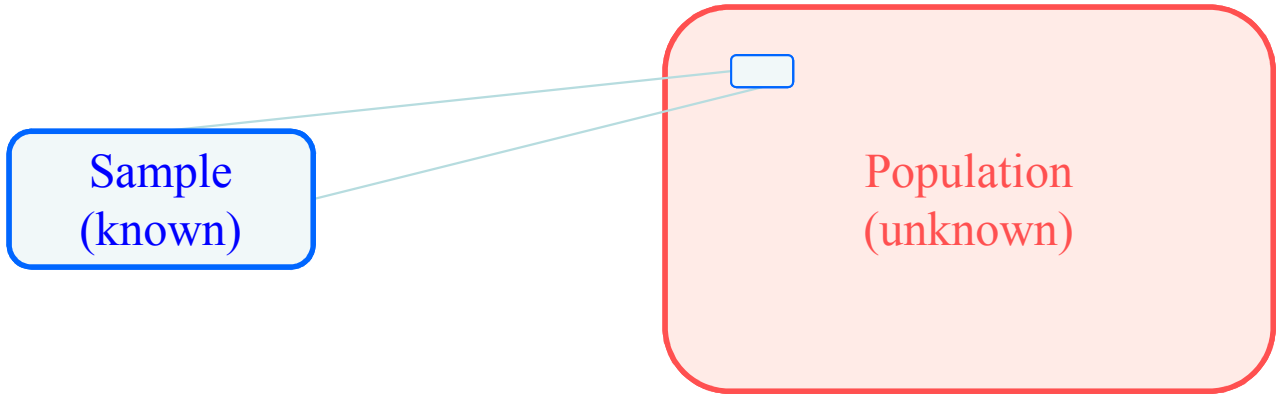
Those values $F(x_1)$, $F(x_2)$, ... ,$F(x_5)$ we can compare to the mid-points of their respective interval on the y-axis: 0.1, 0.3, 0.5, 0.7, and 0.9.

Alternatively, we can compare the original sorted values x_1, x_2, ... ,x_5 to values of $F^{-1}(0.1)$, $F^{-1}(0.3)$, …, $F^{-1}(0.9)$, i.e., the 0.1, 0.3, …, 0.9 quantiles.

The scatter plot of 'theoretical' quantiles (quantiles of mid-points) against the sorted random numbers is called the ***quantile plot*** (or *q-q plot*). Quantile plot is generate by the distribution; for the standard normal this is called *standard normal quantile plot*.

Hypothesis Testing (a view from 30 thousand feet)

Summary: use a sample of data to draw *inferences* about the wider population from which the sample is drawn. In such settings, information about the entire population is, or may be, unknown (and possibly unknowable).



Choose a *test statistic* of interest to describe the population.
For example, if your population is the height of people you may be interested in *average height;* if it is the income you may be interested in *median income*.

Then create a *null hypothesis* (typically a conjecture, i.e., guess[1]) about the population: for example, "the population mean equals…" or "the population variance equals …"

[1] guess could be a commonly accepted fact: for example, the body temperature of a healthy human has a population mean of 98.6º F (37º C).

Classical Statistical Testing

The distribution of the test statistic is known!

For example, the *Sample Mean* of the normal sample has a normal distribution with parameters that are tied to the usually unknown population parameters.

(1) Choose the test statistic for which the distribution is known,

(2) Create a null hypothesis and the alternative hypothesis. These are made about some characteristic of the population related to test statistic in (1),

(3) Compute the test-statistic for the given sample and locate the value on the theoretical distribution from (1). A value in the tail would rarely occur by chance, and the null hypothesis can be *rejected* in favor of the alternative. Otherwise, there is no evidence to reject the null hypothesis – it is *accepted*.

Problem with this approach: many restrictions on *test statistic* and on *population distribution!* It was represented by the classical *t-test*: the hypothesis is about the unknown *population mean*, and the test is devised for the "transformed" *Sample Mean* of the *Normal sample* with unknown population variance.

Statistical Testing via Sampling Distributions

Relies on computing power ("simulations").

(1) Choose a *population property* of interest (e.g., mean, variance, percentile, etc.),

(2) Formulate a null hypothesis about the population property,

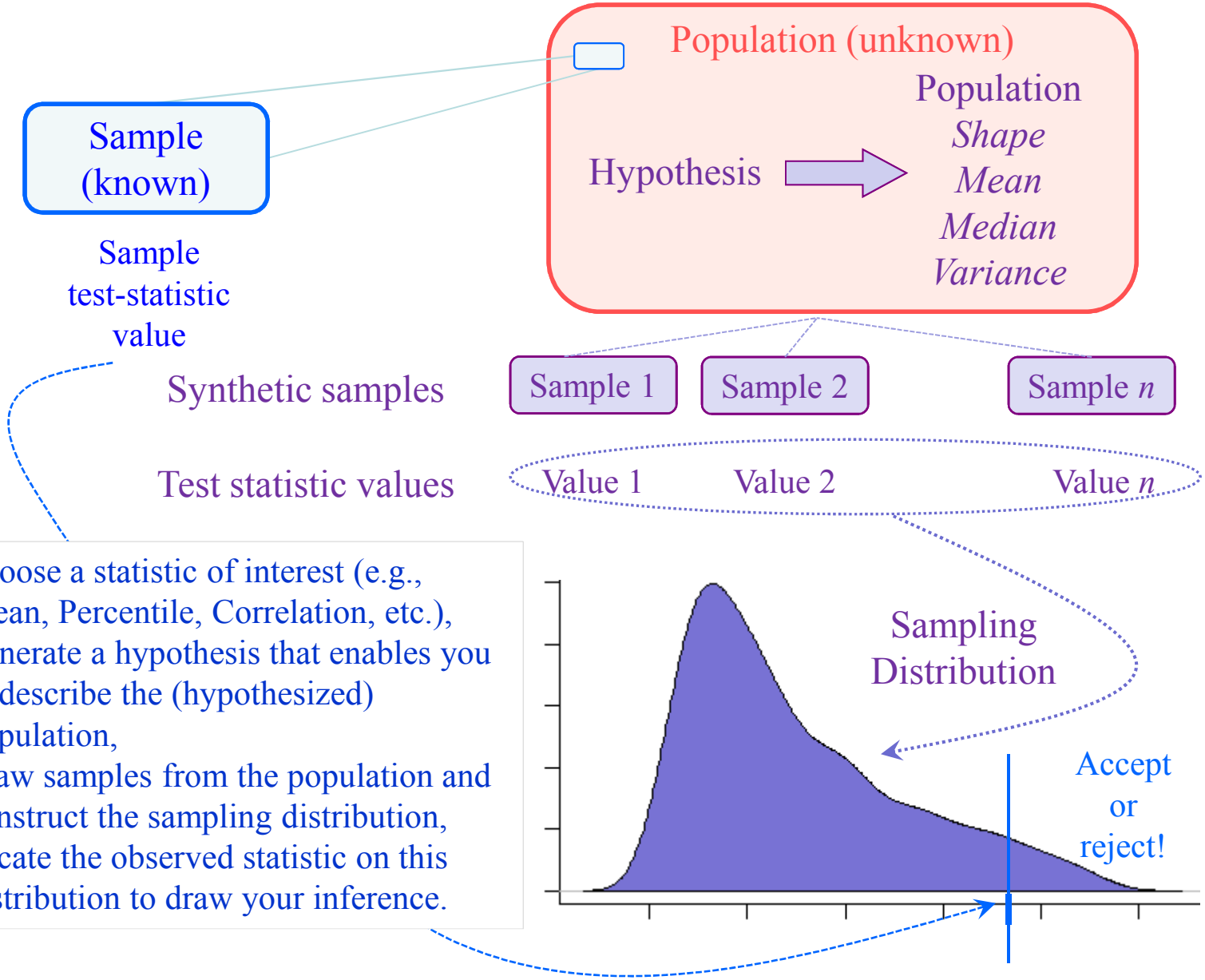
(3) Assume population has some parametric distribution and draw synthetic samples from that distribution **OR** draw synthetic samples by *resampling* from the obtained data (i.e., sample).

(4) Compute the *test stat*istic value (that corresponds to population property of interest) for each of the synthetic samples drawn and create a distribution of these values. This is usually called the *sampling distribution*.

(5) Compute the test statistic for the obtained data and locate the value on the sampling distribution. A value in the tail would rarely occur by chance and the null hypothesis should be rejected. Otherwise, it should be accepted.

With the computing power of *R* you can construct sampling distribution for the test that you are interested in conducting. This approach of creating your own sampling distributions is termed *permutation* (or *randomization*) *testing*. Instead of comparing the actual value of a test statistic to a known distribution (which works under many restrictions), the sampling distribution is generated from the data.

Statistical Testing via Sampling Distributions



1. Choose a statistic of interest (e.g., Mean, Percentile, Correlation, etc.),
2. Generate a hypothesis that enables you to describe the (hypothesized) population,
3. Draw samples from the population and construct the sampling distribution,
4. Locate the observed statistic on this distribution to draw your inference.

Optional

Proof of $E(S^2) = \sigma^2$

Sample Variance of a sample $X_1, X_2, \ldots, X_n$ from a distribution F is an unbiased estimator of the population variance σ^2.

$$E(S_n^2) = E\left(\frac{1}{n-1}\sum_{k=1}^{n}(X_k - \bar{X}_n)^2\right) = \frac{1}{n-1}\sum_{k=1}^{n}E(X_k - \bar{X}_n)^2$$

take an arbitrary k

$$E(X_k - \bar{X}_n)^2 = E(X_k - \mu + \mu - \bar{X}_n)^2 = E\big((X_k - \mu) - (\bar{X}_n - \mu)\big)^2$$

$$= E\big((X_k - \mu)^2 - 2(X_k - \mu)(\bar{X}_n - \mu) + (\bar{X}_n - \mu)^2\big)$$

$$= E(X_k - \mu)^2 - 2E(X_k - \mu)(\bar{X}_n - \mu) + E(\bar{X}_n - \mu)^2$$

$= Var(X_k) = \sigma^2$, since $E(X_k) = \mu$

$= Var(\bar{X}_n) = \frac{\sigma^2}{n}$, since $E(\bar{X}_n) = \mu$

$$E(X_k - \mu)(\bar{X}_n - \mu) = E\left((X_k - \mu)\left(\frac{1}{n}\sum_{i=1}^{n}X_i - \mu\right)\right) = \frac{1}{n}\sum_{i=1}^{n}E\big((X_k - \mu)(X_i - \mu)\big)$$

$$= \frac{1}{n}\sum_{i=1}^{n}Cov(X_k, X_i) = \frac{1}{n}Var(X_k) = \frac{\sigma^2}{n}$$

$Cov(X_k, X_i) = 0$ if $i \neq k$, since X_k and X_i are independent

Proof cont.

Hence: $E(X_k - \bar{X}_n)^2 = \sigma^2 - 2\frac{\sigma^2}{n} + \frac{\sigma^2}{n} = \sigma^2 - \frac{\sigma^2}{n} = \frac{n-1}{n}\sigma^2$, for any k.

$$E(S_n^2) = \frac{1}{n-1}\sum_{k=1}^{n} E(X_k - \bar{X}_n)^2 = \frac{1}{n-1}\sum_{k=1}^{n} \frac{n-1}{n}\sigma^2 = \frac{1}{n}\sum_{k=1}^{n}\sigma^2 = \sigma^2.$$

Upper-tailed CI for population mean calculation

Let Z be the *Transformed Sample Mean* of a normal sample and q_C the C^{th} quantile of the *Standard Normal* distribution. Then

$$Z < q_C \;\Leftrightarrow\; \frac{\overline{X}_n - \mu}{\frac{\sigma}{\sqrt{n}}} < q_C \;\Leftrightarrow\; \overline{X}_n - \mu < \frac{\sigma}{\sqrt{n}} q_C \;\Leftrightarrow\; -\mu < \frac{\sigma}{\sqrt{n}} q_C - \overline{X}_n$$

$$\Leftrightarrow\; \mu > \overline{X}_n - \frac{\sigma}{\sqrt{n}} q_C$$

Hence for $C = 95\%$ and $q_C =$ `qnorm(0.95)`:

$$0.95 = P(Z < q_C) = P\left(\overline{X}_n - \frac{\sigma}{\sqrt{n}} q_C < \mu\right)$$

i.e., with probability 0.95 the population mean μ lies inside $\left(\overline{X}_n - \frac{\sigma}{\sqrt{n}} q_C, +\infty\right)$