

2. Random variable

Normal, Lognormal, Student's t

Specific Learning objectives:

- 2.1. Characterize the Normal Distribution:
State its parameters and main features.
Explain the General 68-95-99% Rule.
Calculate probabilities.
- 2.2. Characterize the Lognormal Distribution:
State its parameters and main features.
Explain its relationship with the Normal Distribution.
- 2.3. Characterize the Student's t Distribution:
State its parameters and main features
Calculate probabilities.

Random variable

- A variable whose value is subject to variations due to chance.
- Can take on a set of possible different values with an associated probability.
- Probability
 - is a numerical quantity between zero and one, and
 - expresses the likely occurrence of an event (the values a random variables takes on).

Examples of random variables:

Discrete scale

Number of cerebro-vascular events diagnosed in the past 5 years
Number of days from study entry to last follow-up visit

Continuous scale

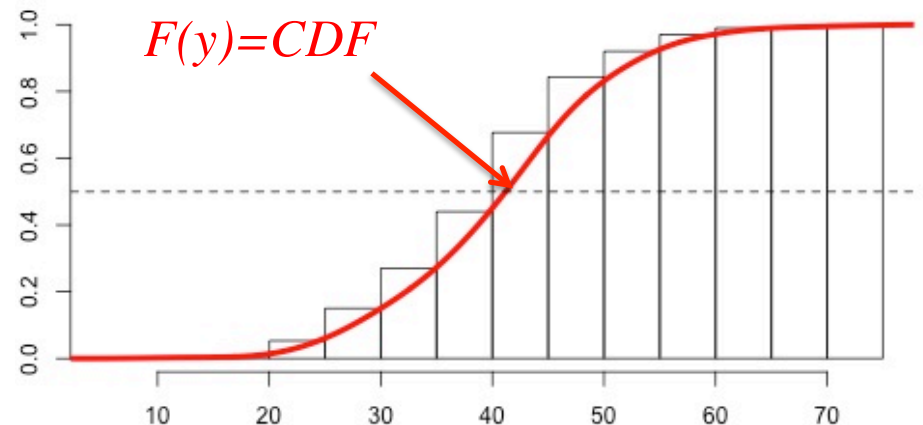
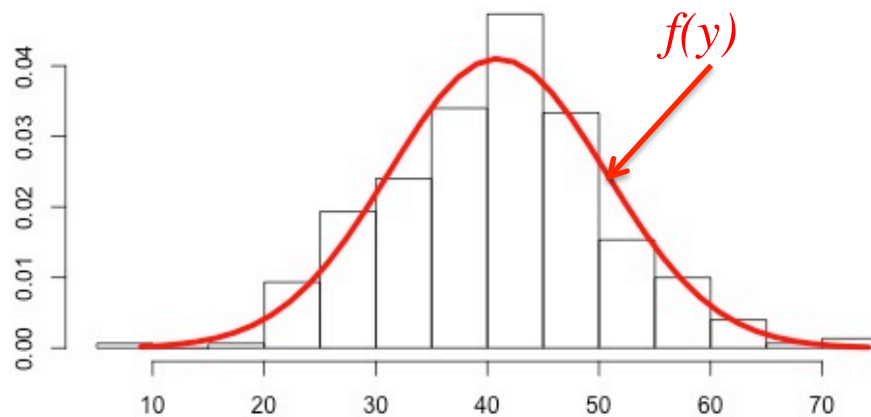
Age
Systolic blood pressure
Concentration of test drug in plasma

} Discussed here regarding
Normal and Lognormal
distributions

Cumulative Distribution Function (CDF) Probability Density Function (PDF) For continuous random variables

$f(y)$ is called the probability density function (PDF) and describes the probability for a random variable to take on set of given values, with two properties:

$$f(y) \geq 0 \quad \text{and} \quad \int_{-\infty}^{\infty} f(y) dy = 1 \quad \text{Area under the curve}$$




Cumulative distribution function (CDF)

$$P(Y \leq a) = \underbrace{\int_{-\infty}^a f(u) du}_{\text{Area under the curve up to } Y=a} = F(a)$$

Area under the curve up to $Y=a$

Expectation and Variance of a Continuous Random Variable

- Used to describe a theoretical distribution, analogous to the sample mean and sample variance.
- Expectation: $E(Y)$, the mean theoretical value, also called mathematical expectation.
 - Expressed with the symbol μ to distinguish it from the arithmetic mean calculated from a sample, \bar{Y} .
- The variance is the measure of spread constructed in terms of $E(Y)$.

$$\mu = E(Y) = \int_{-\infty}^{\infty} y f(y) dy;$$


$$\sigma^2 = Var(Y) = E(Y - \mu)^2.$$

Recall the sample variance definition as an average of the deviation from the mean?

Normal Probability Density Function

The random variable Y is Normally distributed, or $Y \sim N(\mu, \sigma^2)$

The Normal PDF is given by:

$$f(y) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(y - \mu)^2}{2\sigma^2}\right\}, \quad -\infty \leq y \leq \infty$$

Parameters of the Normal Distribution:

$\mu = E(Y)$ specifies the distribution's location

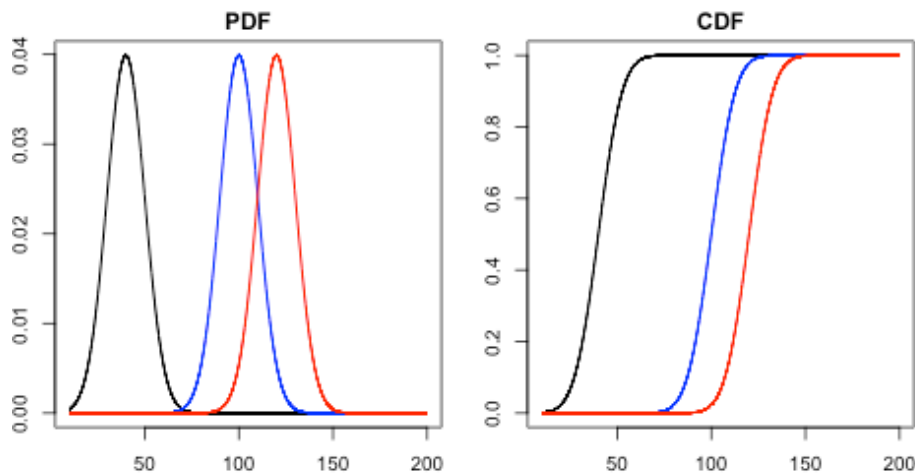
$\sigma^2 = Var(Y)$ specifies the distribution's spread.

On the side: greek letters will be used for parameters and roman letters for sample quantities... e.g. $\mu = E(Y)$ as an example of theoretical quantity or parameter, and \bar{Y} as an example of a sample quantity or statistic.

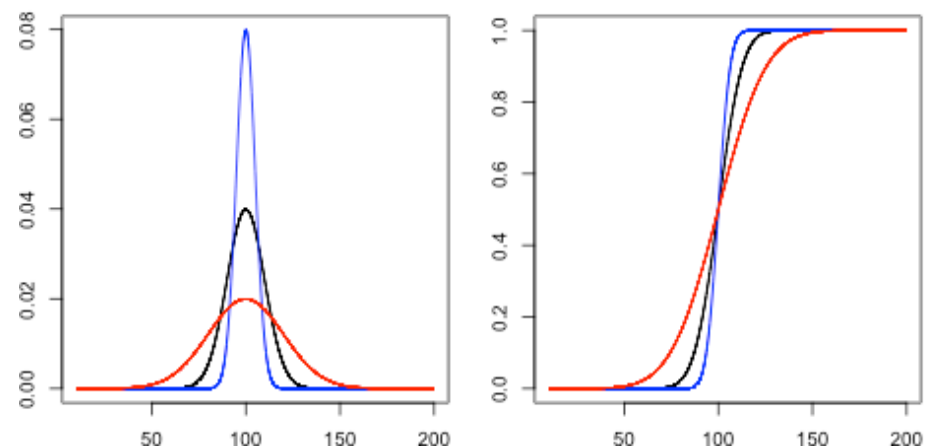
Features of the Normal Distribution

- PDF is symmetric around μ and has a bell-shaped curve
- Y values can go from $-\infty$ to ∞ .
- The highest point of the PDF occurs for μ .
- The shape of the curve is influenced by σ : slimmer curves have smaller σ .
- Data distribution with a bell-curve shape can be approximated with a Normal Distribution.
- $N(0,1)$ is called the Standard Normal Distribution.

$\mu = 40, 100, 120$ $\sigma = 10$



$\mu = 100$, $\sigma = 5, 10, 20$



R functions for the Normal distribution

```
rmnorm(100)                # simulates 100 observations from N(0,1)
rmnorm(100,mean=40,sd=10)  # simulates 100 observations from N(40,100)

pnorm(1)                    # gives the probability below the value of 1 for N(0,1)

dnorm(1)                    # gives the value of the PDF evaluated at 1 for N(0,1)

qnorm(.25)                  # gives the Q1 of the N(0,1)
qnorm(c(.25,.75))           # gives the Q1 and Q3 of the N(0,1)
qnorm(.85)                  # gives the 85th percentile of the N(0,1)
qnorm(.10)                  # gives the 1st decile of the N(0,1)
```

“r” in rmnorm for random

“p” in pnorm for probability

“d” in dnorm for density

“q” in qnorm for quantile

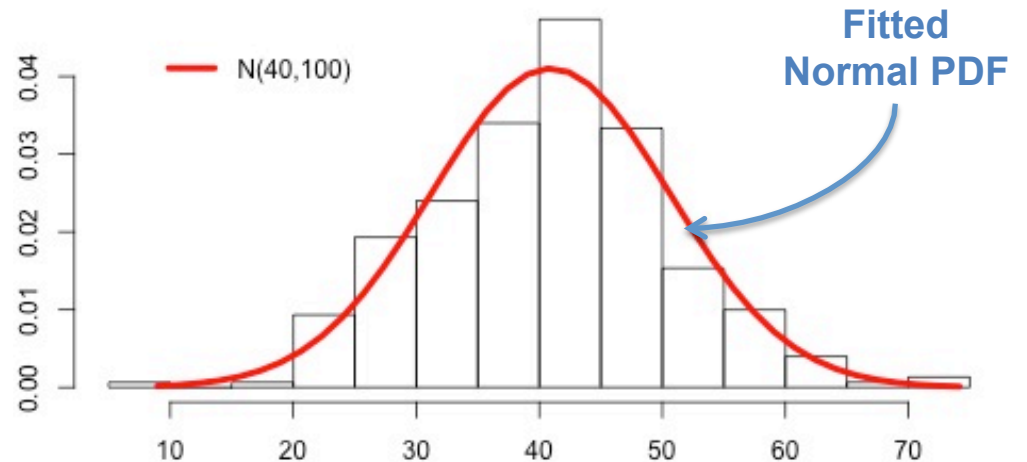
Can specify mean and sd in pnorm, dnorm and qnorm too.

E.g., Y =Hypothetical serum albumin in patients with Primary Biliary Cirrhosis (PBC).
(Based on Altman, 1991)

Scaled histogram vs. $N(40,100)$,
where
 (\bar{X}, s^2) are the sample estimates for (μ, σ^2)
`hist(variable, freq=F)`

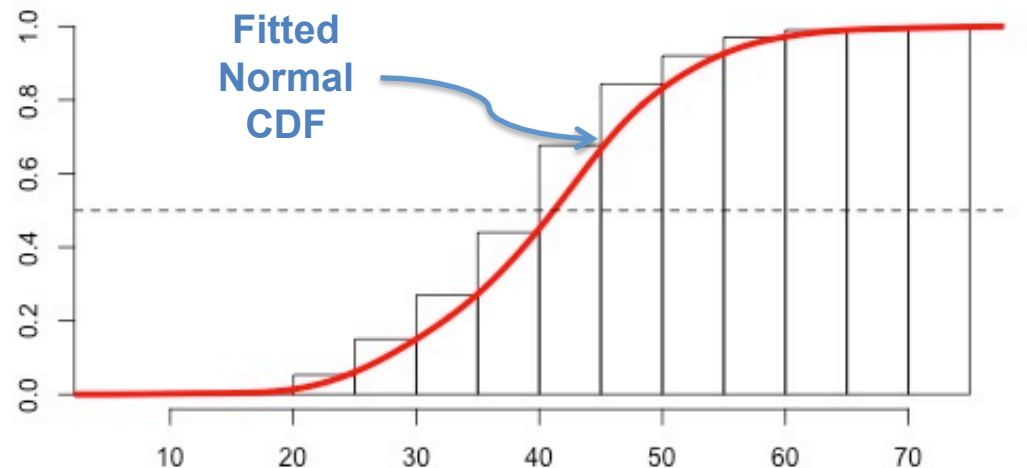
Area under the curve gives a probability:

$$\text{E.g. } P(Y \leq 40) = \int_{-\infty}^{40} f(u) du = 0.5$$



Cumulative histogram vs.
Normal(40,100) CDF:

$N(40,100)$ CDF gives the probability
 $P(Y \leq y)$. E.g. $P(Y \leq 40) = 0.5$



Reasonable assumption: the underlying sample distribution is Normal (40,100)

Calculating probabilities under normality:

Step #1: Standardize the normal variable with mean μ and variance σ^2 :

$$z = \frac{y - \mu}{\sigma}$$

Step #2: Calculate the probability of the standardized variable using the Standard Normal PDF, since

$$Y \sim N(\mu, \sigma^2) \Leftrightarrow Z = \frac{Y - \mu}{\sigma} \sim N(0,1)$$

CASE A: $P(Y \leq y) = P(Z \leq z)$

Example: What is the theoretical proportion of subjects with serum albumin below 43?

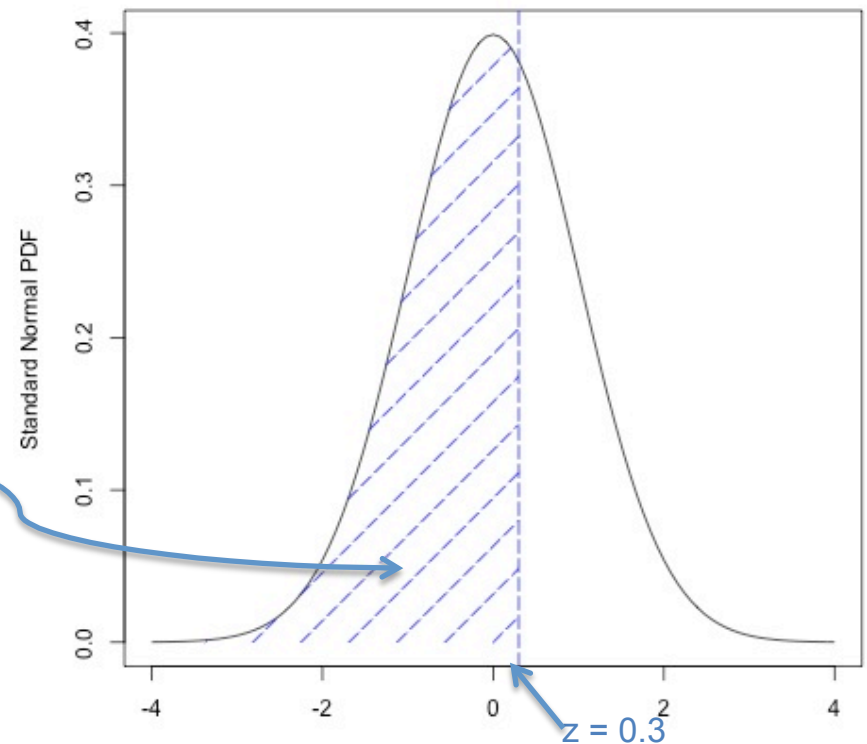
Recall $\mu = 40$ and $\sigma = 10$

$$1. \quad z = \frac{y - \mu}{\sigma} = \frac{43 - 40}{10} = 0.3$$

$$2. \quad P(Z \leq 0.3) = 0.62 = P(Y \leq 43)$$

```
> z <- (43-40)/10  
> pnorm(z)  
[1] 0.6179114
```

```
> pnorm(43,40,10)  
[1] 0.6179114
```



“62% of female individuals have serum albumin lower than 43 g/l”

$$\text{CASE B: } P(y_1 \leq Y < y_2) = P(z_1 \leq Z < z_2)$$

1. Standardize y_1 and y_2

$$z_1 = \frac{y_1 - \mu}{\sigma}, \quad z_2 = \frac{y_2 - \mu}{\sigma}.$$

2. Calculate $P(z_1 \leq Z < z_2) = P(Z \leq z_2) - P(Z \leq z_1)$.

Example: What is the theoretical proportion of subjects with serum albumin between 30 and 50 g/l?

$$z_1 = \frac{30 - 40}{10} = -1, \quad z_2 = \frac{50 - 40}{10} = 1.$$

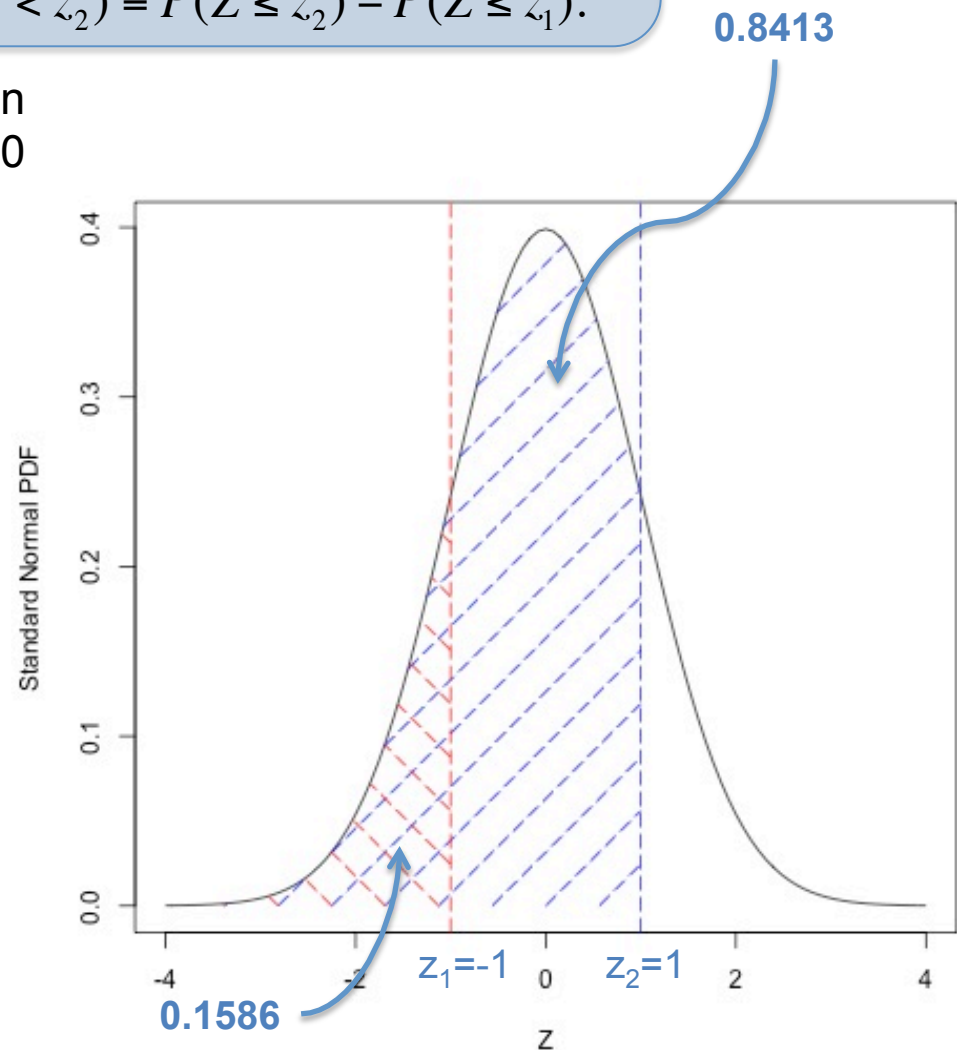
$$P(Z \leq -1) = 0.1586$$

$$P(Z \leq 1) = 0.8413$$

$$P(Z \leq 1) - P(Z \leq -1) = 0.6827 \\ = P(30 \leq Y < 50)$$

```
> z1 <- (30-40)/10
> z2 <- (50-40)/10
> pnorm(z2)-pnorm(z1)
[1] 0.6826895
> pnorm(50,40,10)-pnorm(30,40,10)
[1] 0.6826895
```

“68% of female subjects have serum albumin between 30 and 43 g/l.”



$$\text{CASE B: } P(y_1 \leq Y < y_2) = P(z_1 \leq Z < z_2)$$

1. Standardize y_1 and y_2

$$z_1 = \frac{y_1 - \mu}{\sigma}, \quad z_2 = \frac{y_2 - \mu}{\sigma}.$$

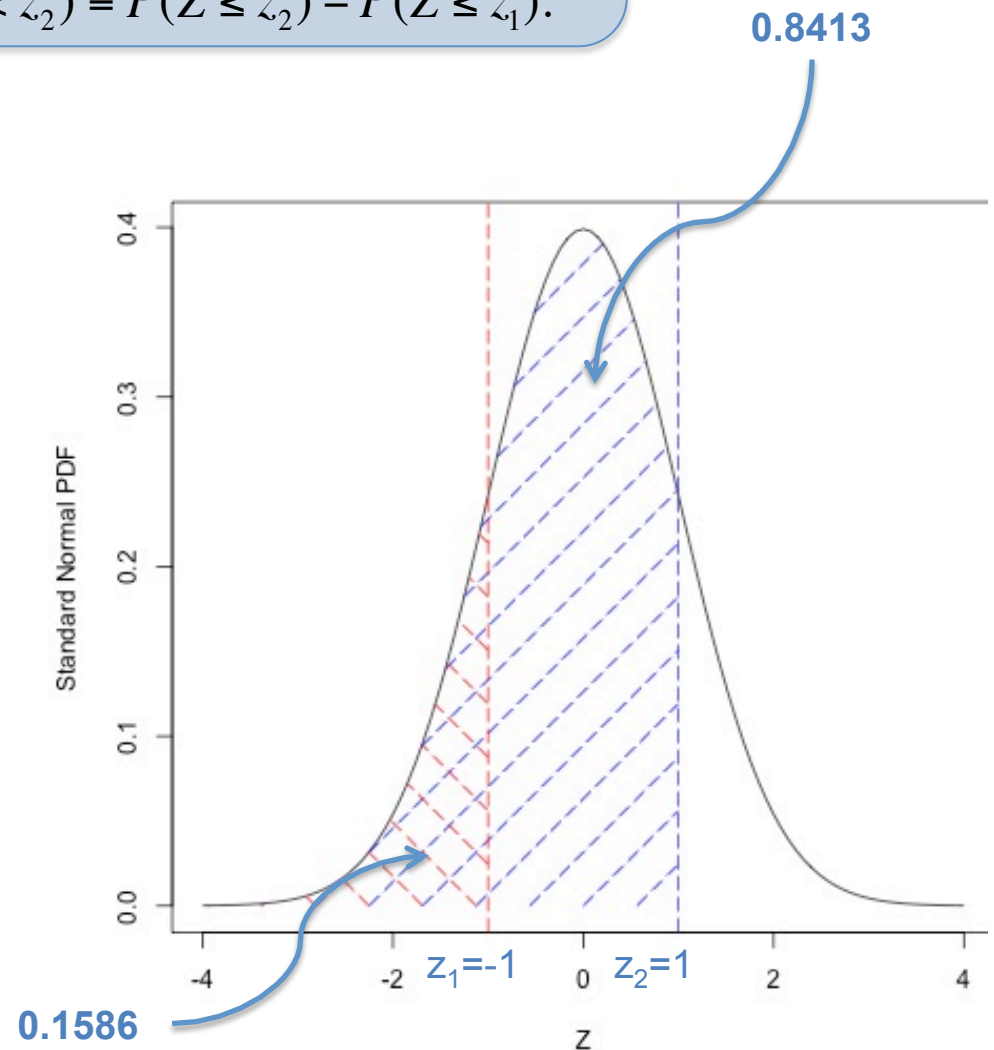
2. Calculate $P(z_1 \leq Z < z_2) = P(Z \leq z_2) - P(Z \leq z_1)$.

Note that this example leads to $z_2=1$ and $z_1=-1$, so can define $z=1$.

We can then simplify the probability statement in step #2 above with

2. Calculate

$$P(-z \leq Z < z) = P(Z \leq z) - P(Z \leq -z).$$



COMPLEMENT OF CASE B when $z_2 = -z_1$: $P(Z \leq -z \text{ or } Z > z) = P(|Z| > z)$

$$\begin{aligned} P(Z \leq z \text{ or } Z > z) &= P(Z \leq -z) + P(Z > z) \\ \text{or, by symmetry,} &= 2P(Z \leq -z) = 2P(Z > z) \\ \text{or by unity of integration,} &= 1 - P(-z < Z \leq z) \end{aligned}$$

What is the theoretical proportion of subjects with serum albumin below 30 and above 50 g/l?

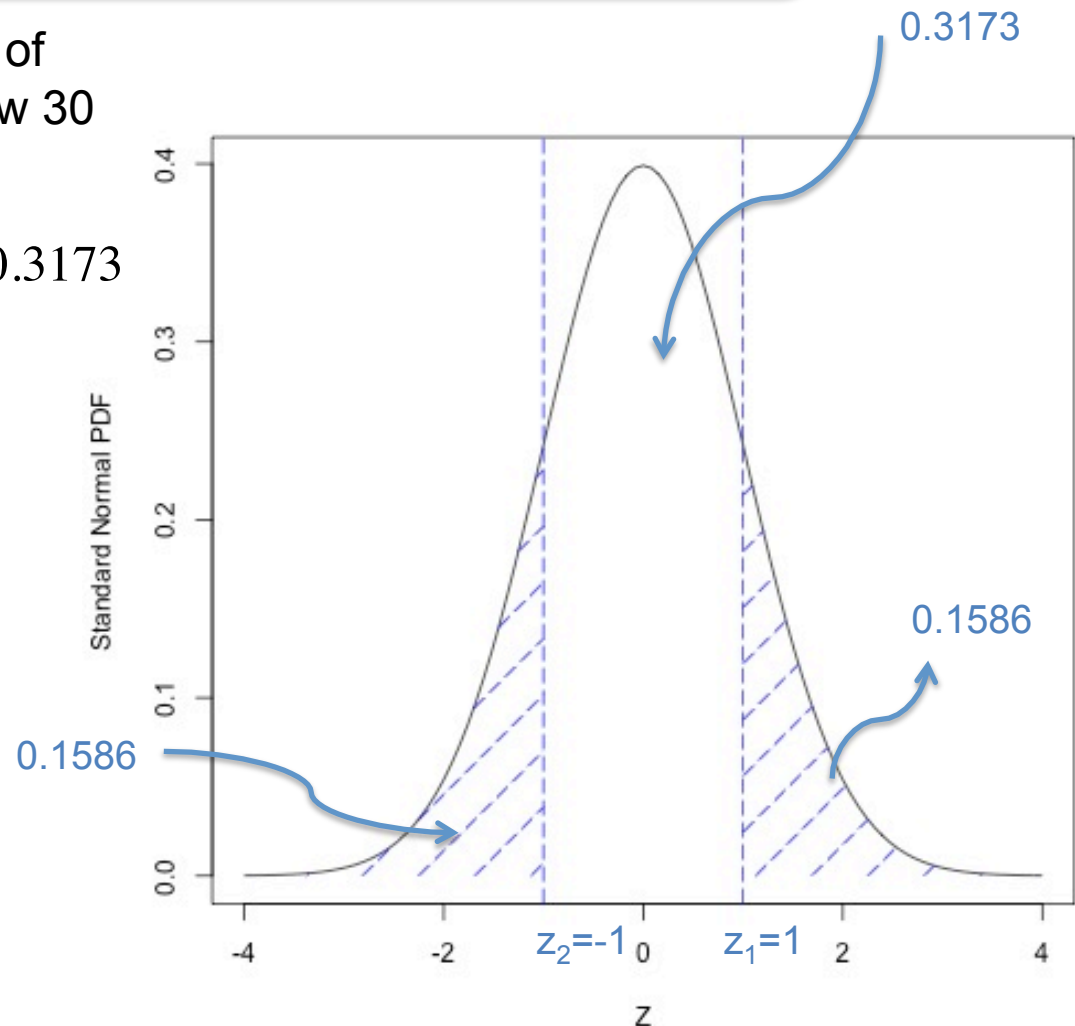
$$1 - P(-1 \leq Z < 1) = 1 - 0.6827 = 0.3173$$

Or, by symmetry,

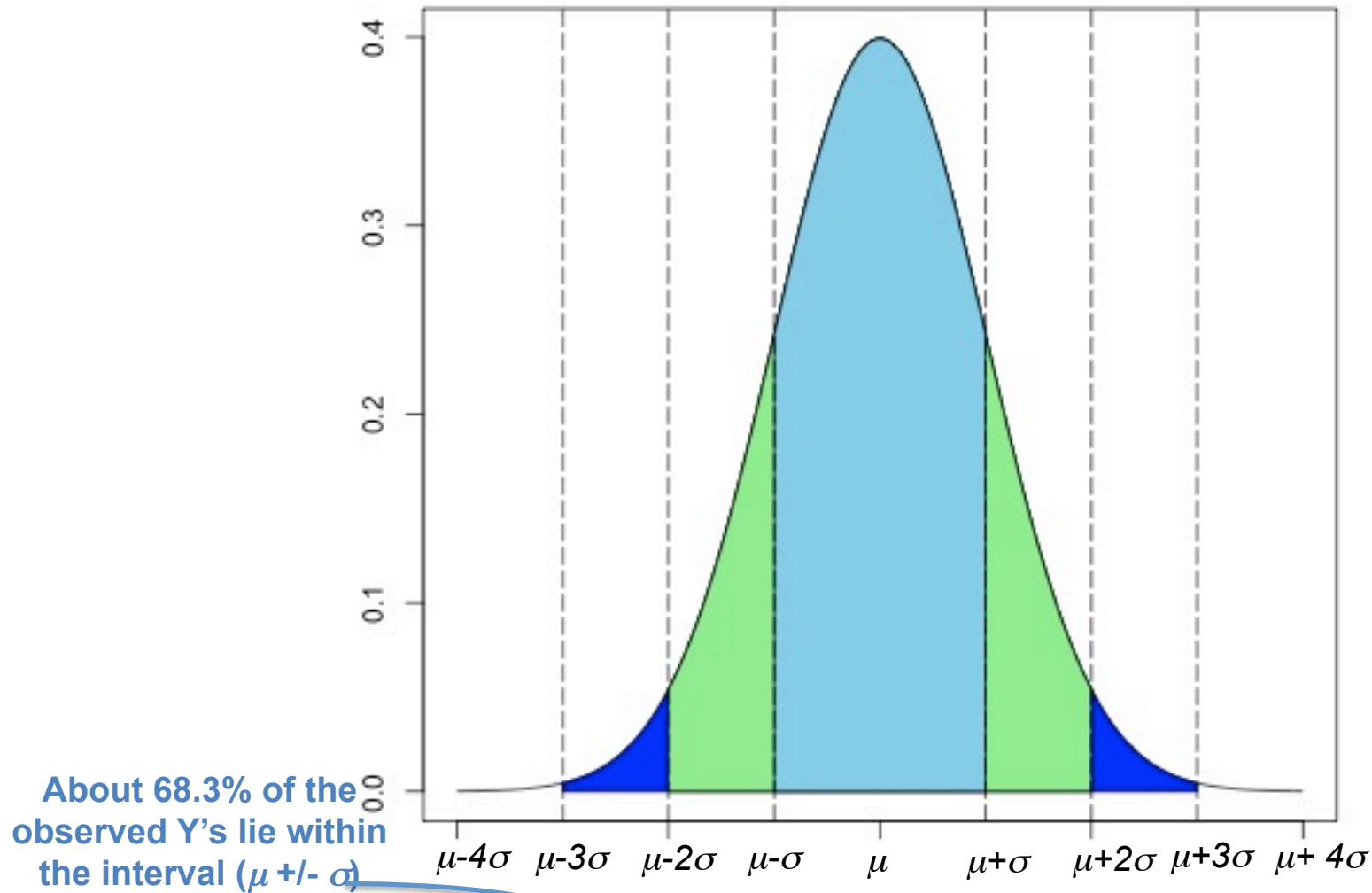
$$2 \times P(Z \leq -1) = 2 \times 0.1586$$

```
> 1 - (pnorm(z2) - pnorm(z1))  
[1] 0.3173105
```

“About 32% of subjects have serum albumin below 30 and above 50 g/l.”



General 68-95-99% Normal Distribution Rule



This rule may roughly apply for sampled data that have a bell-shaped, normal-like distribution.

68.3%

95.4%

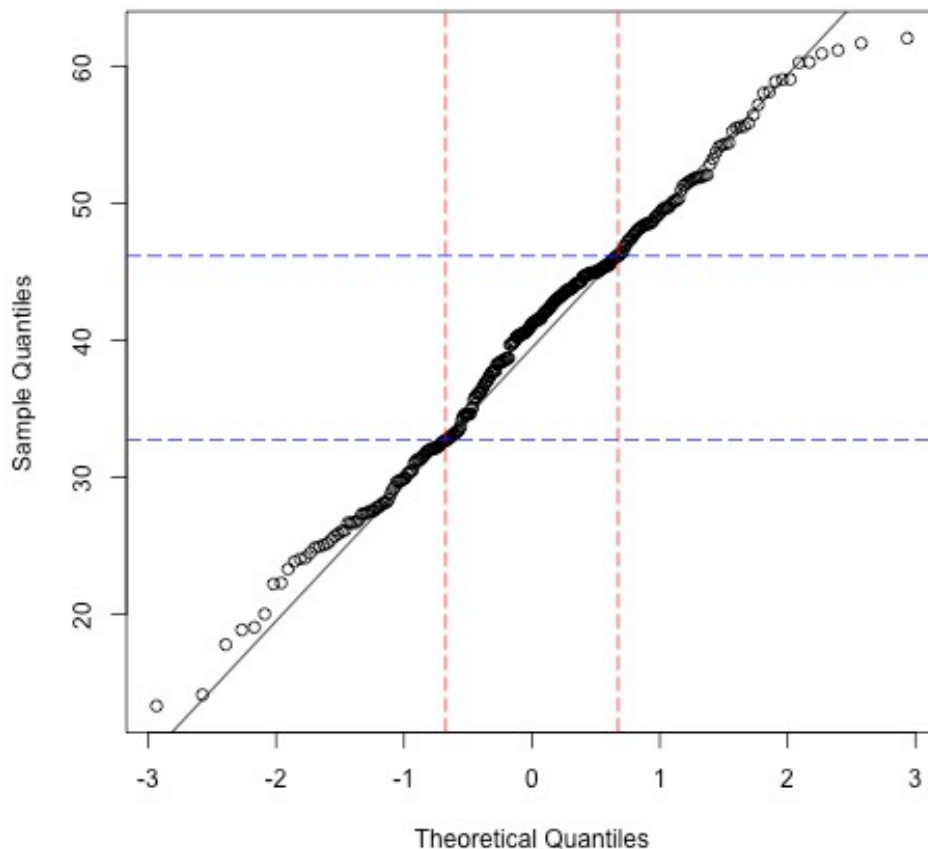
99.7%

For a $N(0,1)$, values >3 are unlikely.

Quantile-Quantile Normal Plot

- Plots the quantiles in the sample vs. theoretical quantiles in $N(0,1)$.
- Should look like a straight line

E.g. Serum Albumin - female data



The line corresponds to the Q1 and Q3 of the sample vs. those of a $N(0,1)$:

```
> quantile(SA.female,c(.25,.75))  
      25%      75%  
32.72200 46.18675  
> qnorm(c(.25,.75))  
[1] -0.6744898  0.6744898
```

Note that an identity line does not make sense since the scales of axes x vs. y may be different)

```
qqnorm(variable) # makes the Quantile-Quantile plot  
qqline(variable) # draws the line connecting Q1's and Q3's
```

Exercise

1. In R, simulate normally distributed $N(120, 8^2)$ values and suppose they represent measurements of systolic blood pressure (SBP) from 50 adult subjects.
2. Assume the underlying distribution of your sample is Normal and estimate μ and σ^2 .
3. Plot a histogram and a Quantile-Quantile plot of the simulated sample against the Standard Normal distribution.
4. Calculate the probability of the following statements under the Normal distribution with the estimated parameters.

Please do so in two ways:

- Standardizing to the $N(0, 1)$
 - Not standardizing but specifying the parameters in the `pnorm` function.
- a) $SBP > 120$,
 - b) $SBP > 139$
 - c) SBP between (110, 130)
5. Find the quantiles that represent the central 68% of the data and plot these as vertical lines in the histogram plotted in (3).

Features of the LogNormal Distribution

Definition of LN
random variable

$$X \sim \text{logNormal} \rightarrow \ln(X) \sim \text{Normal}$$

$$Y \sim \text{Normal} \rightarrow e^Y \sim \text{logNormal}$$

Relationship
between
Normal and LN
parameters

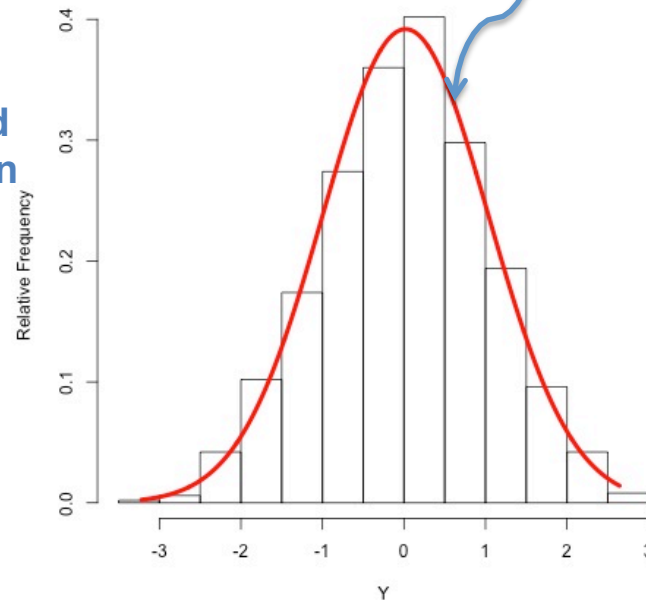
Suppose $\ln(X) \sim N(\mu, \sigma^2)$, then

$X \sim \text{LN}(\mu_L, \sigma_L^2)$ where

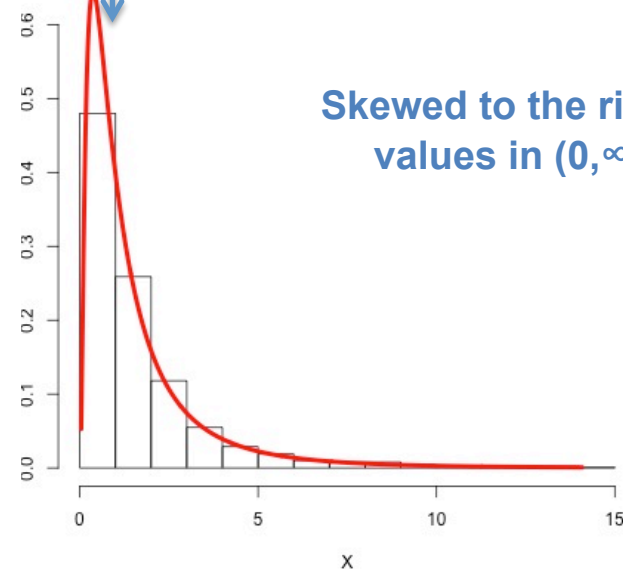
$$\mu_L = E(X) = e^{\mu + \sigma^2/2}, \quad \sigma_L^2 = \text{Var}(X) = (e^{\sigma^2} - 1)e^{2\mu + \sigma^2}$$

... e.g., if $\ln(X) \sim N(0, 1)$ then $X \sim \text{LN}(1.648, 2.161^2)$

Symmetric around
the mean, values in
 $(-\infty, \infty)$



Skewed to the right,
values in $(0, \infty)$



Normal vs. LogNormal

	Normal	LogNormal
Shape	Symmetrical	Skewed to the right
Characteristics	Mean= μ	Mean = $\mu + \sigma^2/2$
	SD= σ	SD = $(\mu + \sigma^2/2) * \sqrt{\exp(\sigma^2) - 1}$
	CV= σ/μ	CV = $\sqrt{\exp(\sigma^2) - 1} \sim (\text{approx}) \sigma^2$
	Median= μ	Median = $\exp(\mu)$
Range of values	$-\infty, \infty$	Strictly greater than 0

Exercise: what would the values of the LogNormal characteristics be for a Normal distribution with $\mu=0$ and $\sigma=1$?

Calculating the CV and Median after log-transformation

Coefficient of variation

For $X = \ln(Y) \sim N(\mu, \sigma^2)$:

$$CV_X = \sigma / \mu$$

For Y, Raw Scale:

$$CV_Y = \sqrt{e^{\sigma^2} - 1} \approx \sigma$$

Median

For $X = \ln(Y) \sim N(\mu, \sigma^2)$:

$$Q2_X = \mu$$

For Y, Raw Scale:

$$Q2_Y = e^\mu.$$

Translating this into usage of a log-transformed data set:

1. Transform Y to $X = \ln(Y)$ so that $X \sim N(\mu, \sigma^2)$.
2. Estimate μ and σ under normality with the sample mean \bar{X} and $sd(X)$
3. Calculate the CV for the raw scale Y, by

$$CV_Y = \sqrt{e^{sd(X)^2} - 1}$$

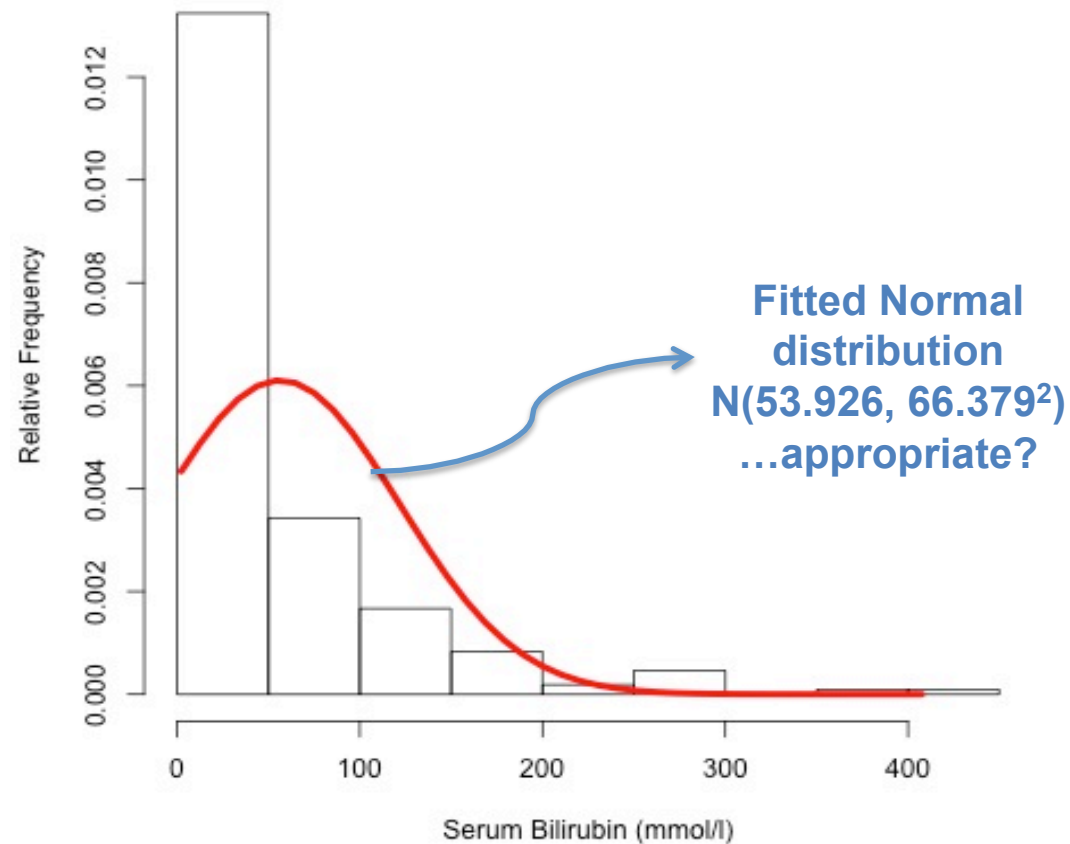
4. Calculate the median for the raw scale by

$$Q2_Y = e^{\bar{X}}$$

Example: Serum Bilirubin (mmol/l) in patients with Primary Biliary Cirrhosis (PBC)
(Based on Altman, 1991)
Serum Bilirubin has been used as a marker for liver disease.

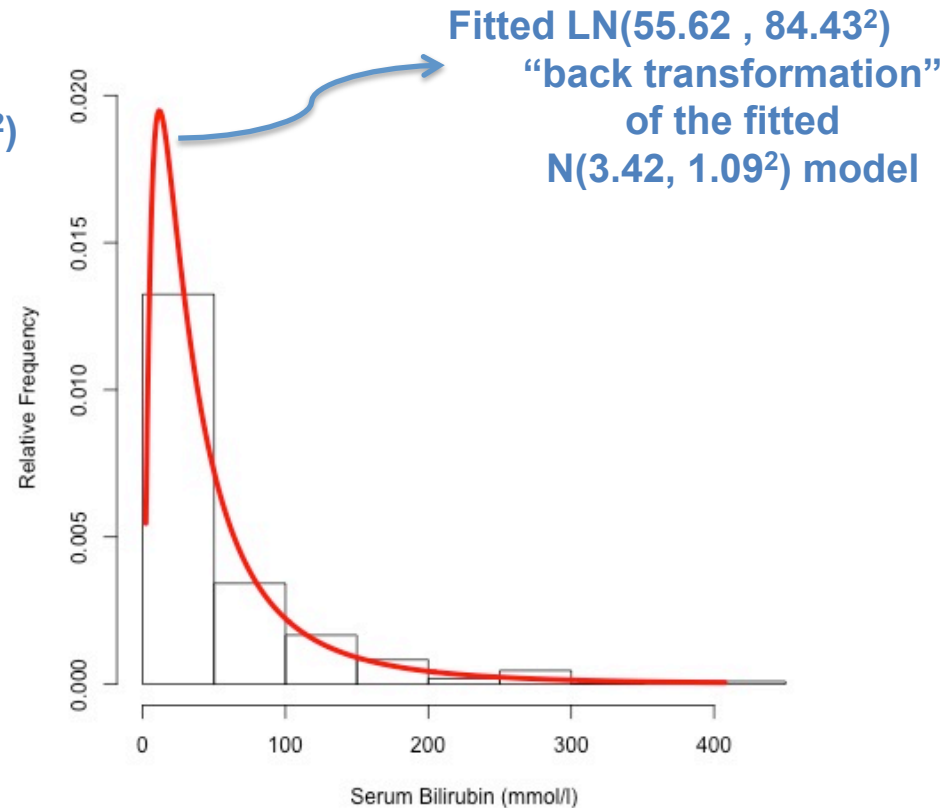
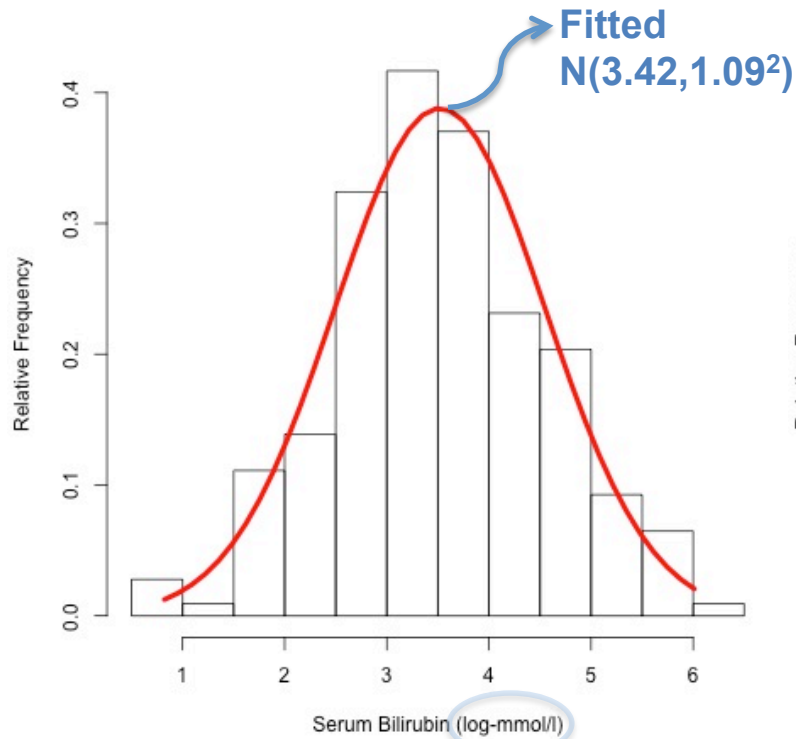
Assumption : $X = \text{Serum Bilirubin} \sim N(53.926, 66.379^2)$

Sample mean and sd^2 of X



Sample mean and sd^2 of $\ln(X)$

Assumption : $\ln(X) = \ln(\text{Serum Bilirubin}) \sim N(3.421, 1.093^2)$



- In log units, 95.4% of the distribution will be expected to be between

$$(\mu - 2\sigma, \mu + 2\sigma) = (1.234, 5.607)$$

- In original units, 95.4% of the distribution will be expected to be between

$$\exp\{ (1.234, 5.607) \} = (3.437, 272.457)$$

- The antilog of the mean of the transformed data is $\exp\{3.421\} = 30.60$

Student's t Distribution

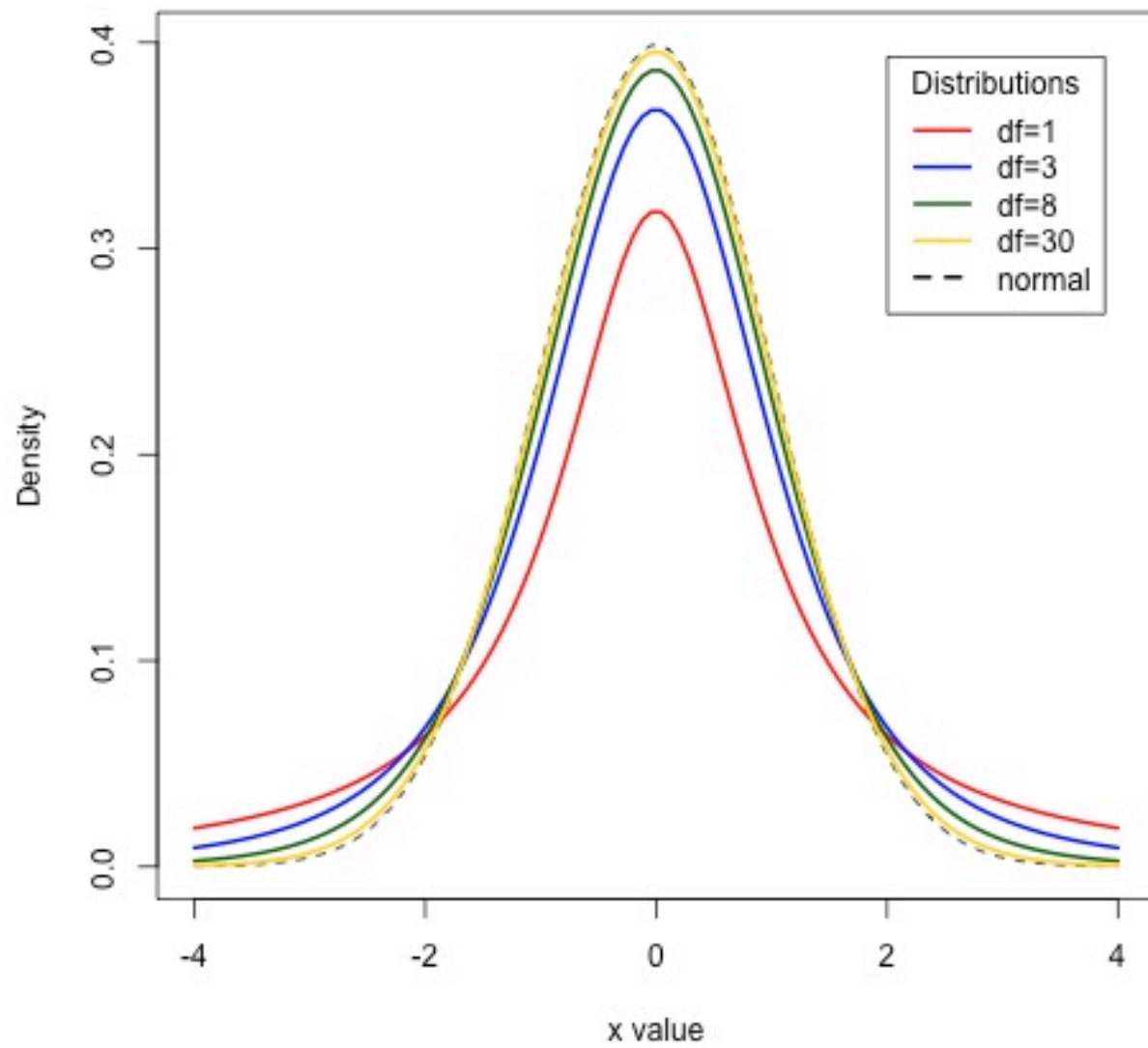
Main features:

- Continuous probability distribution, symmetric and bell-shaped.
- Has heavier tails than the Normal distribution.
- Used when sample size is small and population standard deviation is unknown.
- The larger the sample, the more similar to a Normal distribution.

Features of the Student's t Distribution

- Parameter: degrees of freedom (df).
- Some applications:
 - Test the difference between two sample means.
 - Construct confidence intervals for the difference between two population means.
 - Test significance of linear regression analysis coefficients.

Comparison of t Distributions



Comparison of probabilities under Standard Normal and Student's t distributions

Example: Serum albumin (g/l), with $\mu = 40$ and $\sigma = 10$.

Serum Albumin	N(40,100)	t, df=9	t, df=100	t, df=2000
< 43	0.618	0.614	0.618	0.618
[30,50]	0.683	0.657	0.680	0.683
<30, >50	0.317	0.343	0.320	0.317

```
> pt(z,df=9)
[1] 0.6145046
> pt(z2,df=9)-pt(z1,df=9)
[1] 0.6565636
> 1-(pt(z2,df=9)-pt(z1,df=9))
[1] 0.3434364
```

Other related R functions analogous to those for the Normal distribution

qt() # for quantiles

d(t) # for density function values

rt() # for simulation

Theoretical Quantiles for the Student's t Distribution

The notation used for a theoretical quantile is

$$t_{p,df}$$

E.g.,

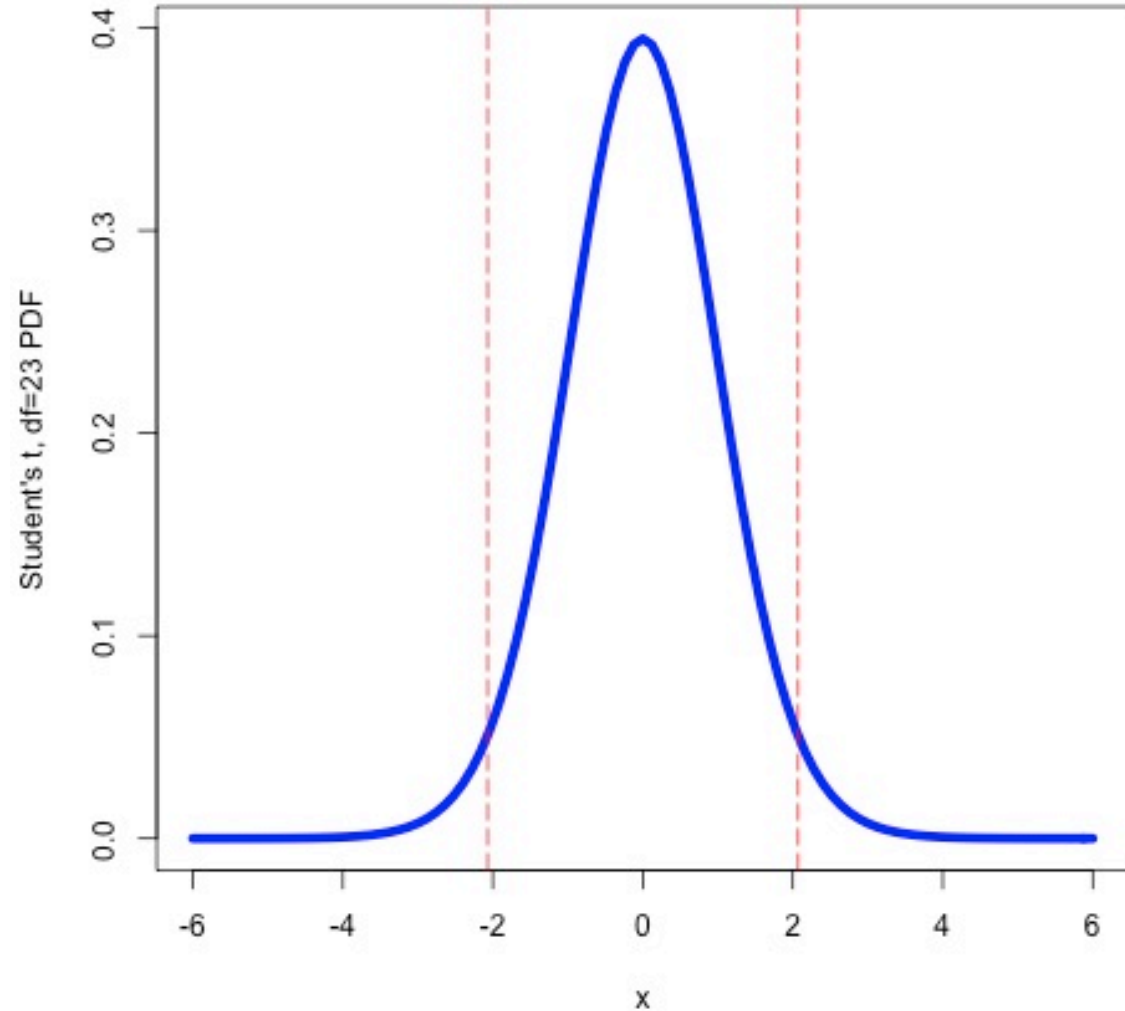
$$t_{0.025,df=23} = -2.068$$

$$t_{0.975,df=23} = 2.068$$

Note that by symmetry,

$$-t_{0.975,df=23} = t_{0.025,df=23}$$

```
> qt(.025,df=23)
[1] -2.068658
> qt(.975,df=23)
[1] 2.068658
```



3. Sampling distribution and inferential statements under Normality

Specific learning objectives:

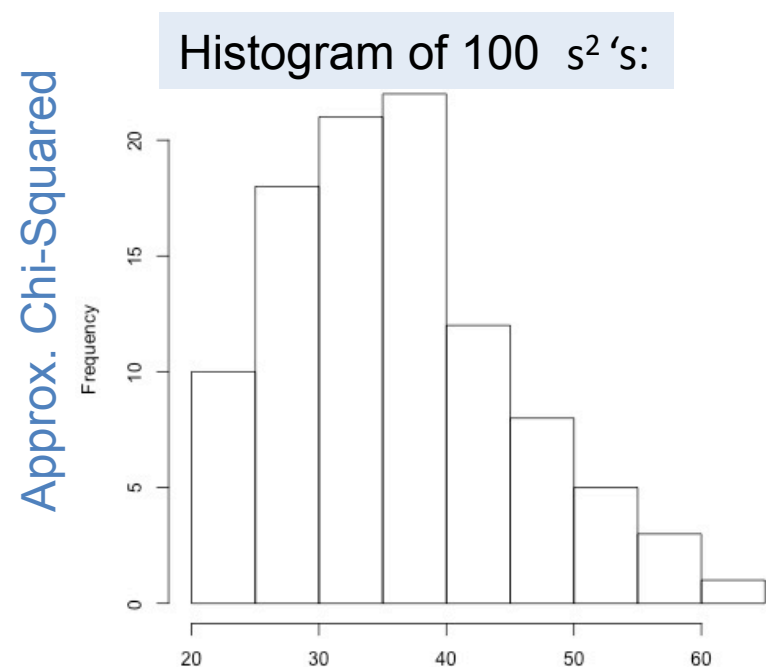
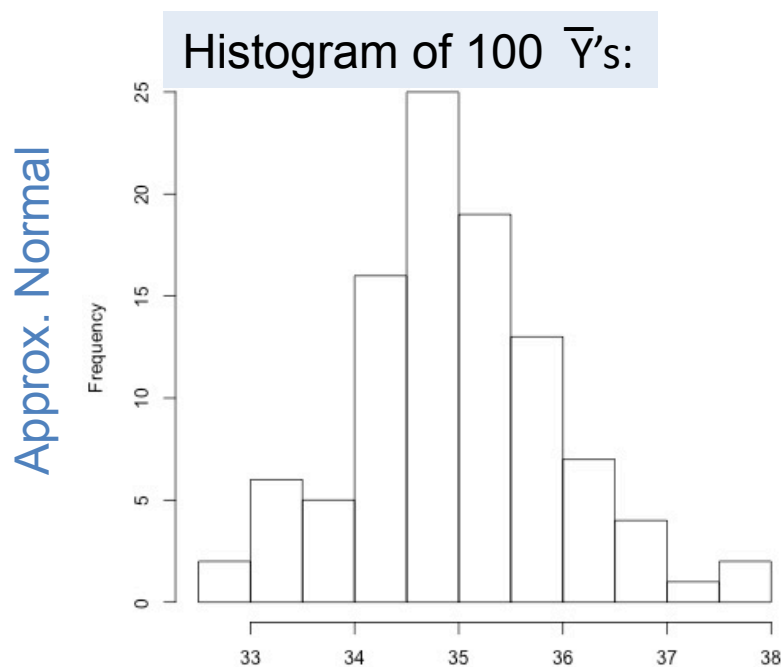
3.1. State, regarding sampling distributions: definition, usage, the sampling distribution of the mean statistic under normality.

3.2. Calculate Confidence Intervals (CI's) for the population mean, explain their interpretation in terms of confidence level.

3.3. Calculate Hypothesis Tests involving population means, describe the three main components (hypothesis statement, test statistic, decision rule). Explain their interpretation in terms of the significance level and p-value.

Sampling variation

- Suppose that the temperature of a random sample of 35 subjects is measured and assume it is Normally distributed.
- We can estimate (μ, σ^2) with the sample mean and sd: (\bar{Y}, s^2) .
- Suppose we can take a second sample of 35 subjects and measure their temperature. We again obtain the estimates (\bar{Y}, s^2)
- Doing this procedure 100 times, we'll have 100 \bar{Y} 's and 100 s^2 's.



Sampling distribution and SE of the mean

- The sampling distribution of the sample mean \bar{Y} is:

$$\bar{Y} \sim N(\mu, \sigma^2 / n)$$

- The population standard error of the mean or $SE(\bar{Y})$ is:

$$\sigma_{\bar{Y}} = \frac{\sigma}{\sqrt{n}}$$

- **Central Limit Theorem** : If n is large (i.e., $n > 30$), regardless of the underlying distribution of the Y 's, the sampling distribution of the sample mean \bar{Y} is Normally distributed with mean μ and SE σ/\sqrt{n} .

Sampling distributions in general

Sampling distribution in general: the **probability distribution** of a given **statistic** (which is based on a random **sample**, e.g. \bar{Y}).

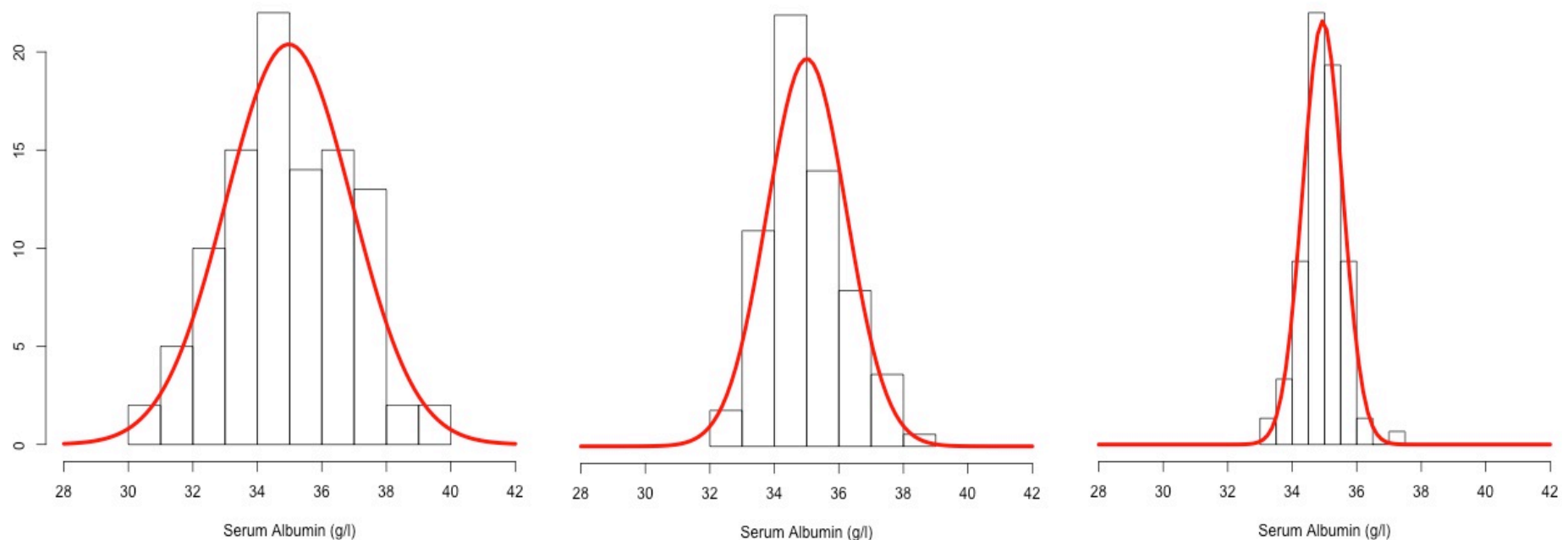
Sampling distributions are used to:

1. Assess how close the value of the statistic is likely to be to the -unknown- population parameter (e.g. sample mean vs. population mean) by estimating its sampling variation.
2. We can do this with a single sample, using probability theory and the CLT... i.e., no need to take 100 samples!
3. Every sample statistic, whether it is mean \bar{Y} , or variance s^2 , has a SE or measure of uncertainty associated with it.

This can be empirically verified via simulations.

1. Simulate $Y = \text{serum albumin (g/l)} \sim N(\mu=35, \sigma^2=36)$ for $n=10$ subjects.
2. Calculate the sample mean and sd.
3. Repeat (1) and (2) 100 times for $n=10, 25, 100$.

The true distribution of the sample mean is $\bar{Y} \sim N(\mu, \sigma^2/n) = N(35, 36/n)$



Dispersion decreases with larger $n \iff$ lower uncertainty of the estimate.

Example of R code for simulating 100 samples from a Normal distribution

`numeric()`, for loop, `par(mfrow(,))`

samples of size 10

Preamble	{	<code>means.10 <- numeric()</code>	<code># empty object to store means</code>
		<code>vars.10 <- numeric()</code>	<code># empty object to store vars</code>
Body	{	<code>for (i in 1:100){</code>	<code># start of "for" loop for 1-100</code>
			<code># iterations</code>
		<code>sample.10 <- rnorm(10,35,6)</code>	<code># simulation of one sample size 10</code>
		<code>means.10[i] <- mean(sample.10)</code>	<code># storing i-th mean in i-th place</code>
		<code>vars.10[i] <- var(sample.10)</code>	<code># storing i-th sd in i-th place</code>
		<code>}</code>	<code># end of "for" loop</code>
		<code>hist(means.10,freq=F,xlim=c(value1,value2))</code>	

Can add `means.25`, `vars.25`, `means.100`, `vars.100`, etc.

Type `par(mfrow=c(1,3,))` to plot the three histograms (for `n=10,25,100`) in one column

Confidence interval (CI)

- A range of values which we can be confident includes the true parameter value.
- Covers a proportion of the sampling distribution of the statistic of interest.

E.g., Given that $\bar{Y} \sim N(\mu, \sigma^2 / n)$, by the 68 - 95 - 99% Normal Distribution Rule we can say :

A 68.3% CI for μ is: $(\bar{Y} - 1 \text{ SE}(\bar{Y}), \bar{Y} + 1 \text{ SE}(\bar{Y}))$

Or: A 95.4% CI for μ is: $(\bar{Y} - 2 \text{ SE}(\bar{Y}), \bar{Y} + 2 \text{ SE}(\bar{Y}))$

Or: A 99.7% CI for μ is: $(\bar{Y} - 3 \text{ SE}(\bar{Y}), \bar{Y} + 3 \text{ SE}(\bar{Y}))$

A 95% CI for μ is: $(\bar{Y} - 1.96 \text{ SE}(\bar{Y}), \bar{Y} + 1.96 \text{ SE}(\bar{Y}))$

Confidence level ↑

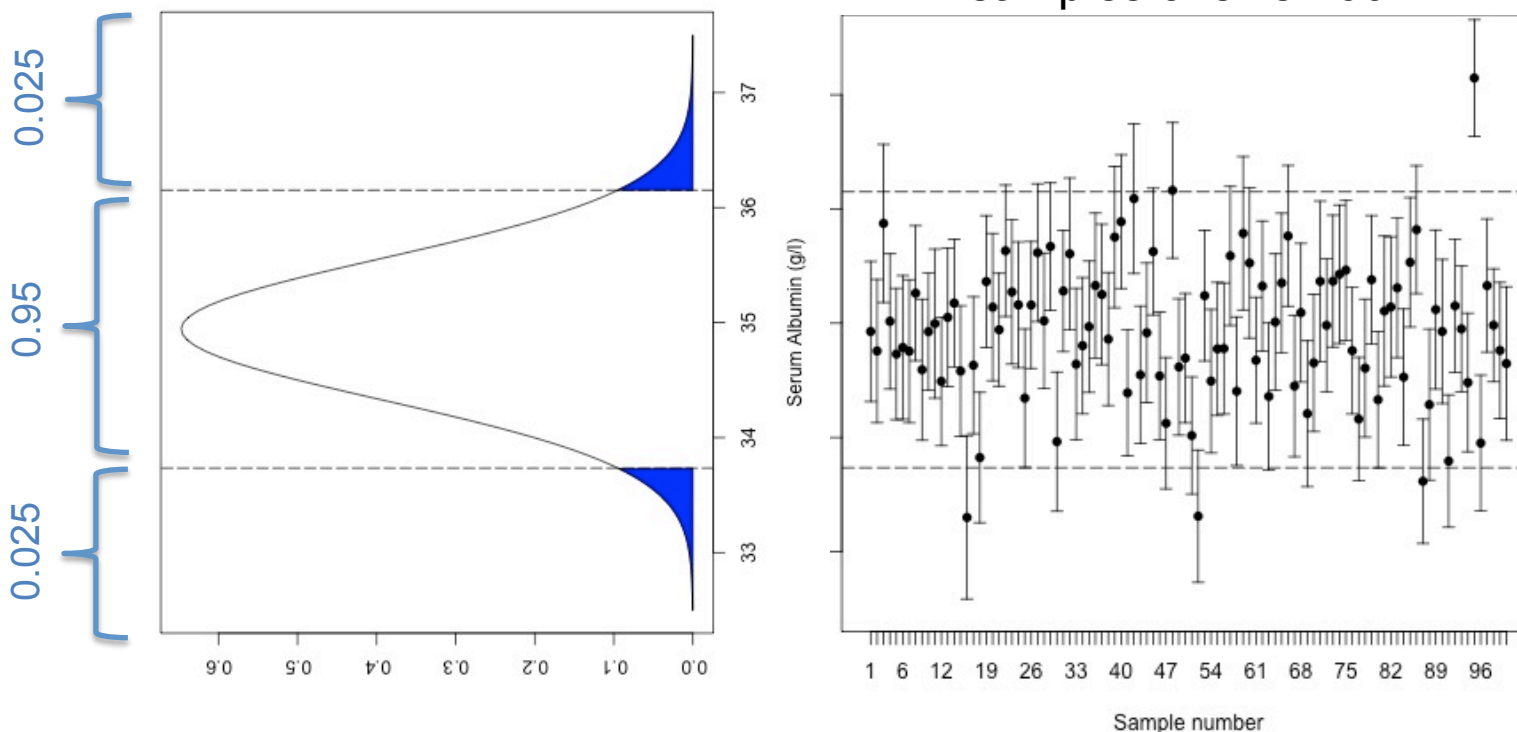
Will not include μ 5% of the time

Quantiles for the 2.5-th and 97.5-th percentiles.... $\text{qnorm}(.025)=-1.96$, $\text{qnorm}(0.975)=1.96$

The 95% CI for the sample mean is the range of values which contains the true population mean with probability 0.95.

E.g. Generated Serum Albumin data, $\mu=35$, $n=100$

CI's for mean Serum Albumin constructed from 100 random samples of size 100.



Horizontal lines show the range within 95% of the sample means are expected to fall:
95% CI for μ : (33.73, 36.15)

(1- α)x100% CI for the mean using the
Standard Normal and Student's t Distributions

Standard Normal Distribution: $(\bar{Y} - z_{1-\alpha/2}SE(\bar{Y}) , \bar{Y} + z_{1-\alpha/2}SE(\bar{Y}))$

E.g. for $\alpha = 0.05$, $z_{0.975} = 1.96$: $(\bar{Y} - 1.96 SE(\bar{Y}) , \bar{Y} + 1.96 SE(\bar{Y}))$

Z <- qnorm

qnorm gives the 0.975th of the N(0,1)

Students' t Distribution: $(\bar{Y} - t_{1-\alpha/2,df}SE(\bar{Y}) , \bar{Y} + t_{1-\alpha/2,df}SE(\bar{Y}))$

E.g. for $\alpha = 0.05$ & $df = 10$,

$t_{0.975,df=10} = 2.23$: $(\bar{Y} - 2.23 SE(\bar{Y}) , \bar{Y} + 2.23 SE(\bar{Y}))$

```
> qnorm(.975)  
[1] 1.959964
```

```
> qt(0.975,df=10)  
[1] 2.228139
```

Exercise Student's t Distribution

1. In R, download the data set “EnergyIntake_tDist.csv” from the LEARN site. This data set consists of the average energy intake (kJ) of 11 healthy women aged 22-30 years.
(Altman, 1991)
2. Calculate the 95% confidence interval for the mean average energy intake, assuming that these data can be approximated by a Normal distribution.
3. Based on the 95% CI, is the average woman in this sample having the recommended daily intake of 7725 kJ? On average, is there a deficit or a surplus in energy intake in this sample?

Hypothesis test

- Statistical method in which data (evidence) is used to choose between two decisions.
 - E.g. Treatment does not increase plasma glucose vs. Treatment does increase plasma glucose.
- Based on “proof by contradiction”
- Process of hypothesis testing:
 1. State null hypothesis: aimed to be disproved by collecting evidence that contradicts it, in favor of the alternative hypothesis.
 2. State alternative hypothesis: to be concluded as a result of the research.
 3. Decide on a test statistic that if observed, could sufficiently contradict the null.
 4. Calculate the probability that could have been obtained from the observed statistic *if the null was true*: the smaller, the more untenable is the null hypothesis.

Types of Errors

Type I: $\alpha = P(\text{Reject } H_0 \mid H_0 \text{ is true}) = \text{significance level}$

Type II: $\beta = P(\text{Do not reject } H_0 \mid H_0 \text{ is false})$

← Main focus in course

1- α and 1- β are the probabilities of drawing the correct conclusion.

	H_0 true	H_0 false
Reject H_0	α	1- β
Do not reject H_0	1- α	β

← “Power”, the larger the better

Research studies must be planned to minimize these errors, based on practical implications.

An extreme analogy...

“The accused is assumed innocent until proven guilty”

H_0 : subject is innocent vs. H_1 : subject is guilty

$\alpha = P(\text{Penalize} \mid \text{person is innocent})$ and

$\beta = P(\text{Not penalize} \mid \text{person is guilty})$

Recall: H_0 is aimed to be disproved by collecting evidence that contradicts it, in favor of the alternative hypothesis.

Examples of hypothesis tests for the population mean (Parametric methods)

- One sample tests

H_0 : the population mean μ is equal to a hypothesized value μ_0 .

- Two (independent) samples tests

H_0 : the two samples come from populations with the same means.

- Paired tests

H_0 : the true means of a single sample under two different conditions are equal.

- Can be simplified to one sample tests by analyzing the difference in response values as primary variable.

The Student's t distribution is used as an approx. to the Normal when variance is unknown and estimated.

- Analysis of Variance (ANOVA) tests: comparison of the population mean of more than two groups.

Hypothesis test for the population mean, one sample test.

Null: the population mean μ is equal to a hypothesized value μ_0 .

Two sided alternative : μ is not equal to μ_0

$$H_0 : \mu = \mu_0 \quad \text{vs.} \quad H_1 : \mu \neq \mu_0$$

Equivalently, $H_0 : \mu - \mu_0 = 0 \quad \text{vs.} \quad H_1 : \mu - \mu_0 \neq 0$

The alternative here is two sided: if true,
implies the following possibilities

$$\mu < \mu_0 \quad \text{or} \quad \mu > \mu_0$$

(equivalently, $\mu - \mu_0 > 0 \quad \text{or} \quad \mu - \mu_0 < 0$)

One sided alternative: μ is greater to μ_0 (or μ is less than μ_0)

$$H_0 : \mu = \mu_0 \quad \text{vs.} \quad H_1 : \mu > \mu_0 \quad (\text{or } \mu < \mu_0)$$

One sided is not as common and left out of the course contents.
Both hypotheses have to cover all possible values of the quantity of μ

Test statistic

- A function of the data that is used to disprove H_0 .
- Its value is compared with the known distribution of what we expect when the null is true.

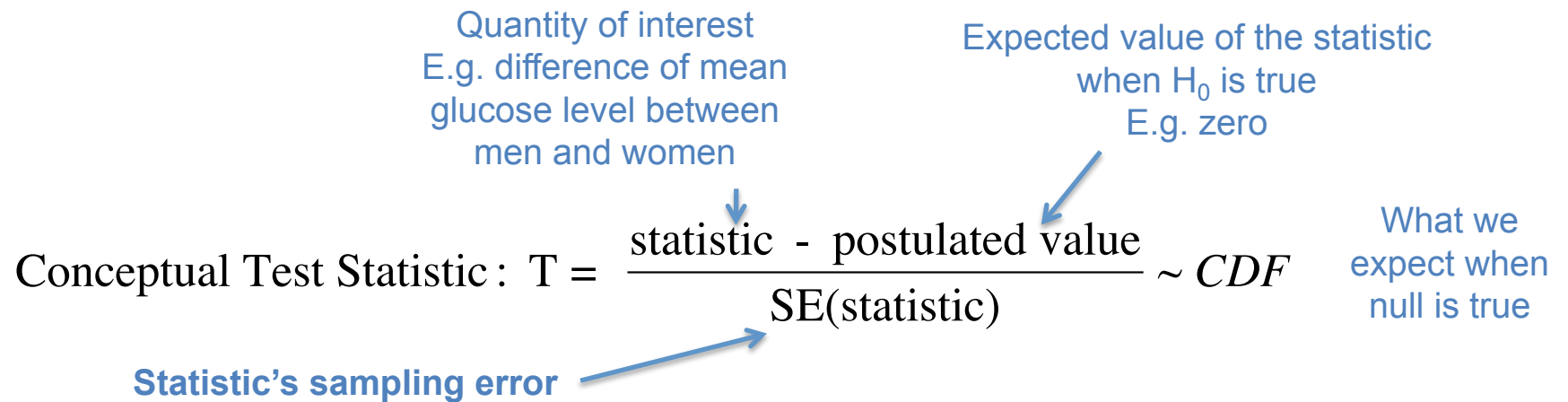
Quantity of interest
E.g. difference of mean
glucose level between
men and women

Expected value of the statistic
when H_0 is true
E.g. zero

What we
expect when
null is true

Conceptual Test Statistic: $T = \frac{\text{statistic} - \text{postulated value}}{\text{SE}(\text{statistic})} \sim CDF$

Statistic's sampling error



Observed Test Statistic: $T_0 = \frac{\text{Obs. statistic} - \text{postulated value}}{\text{obs. SE}(\text{statistic})}$

T_0 is observed, based on the
sample at hand

How likely is T to occur under the assumption that H_0 is true?
What is the probability for T to have an observed value at least as extreme as T_0 ?

Decision Rule, p-value approach

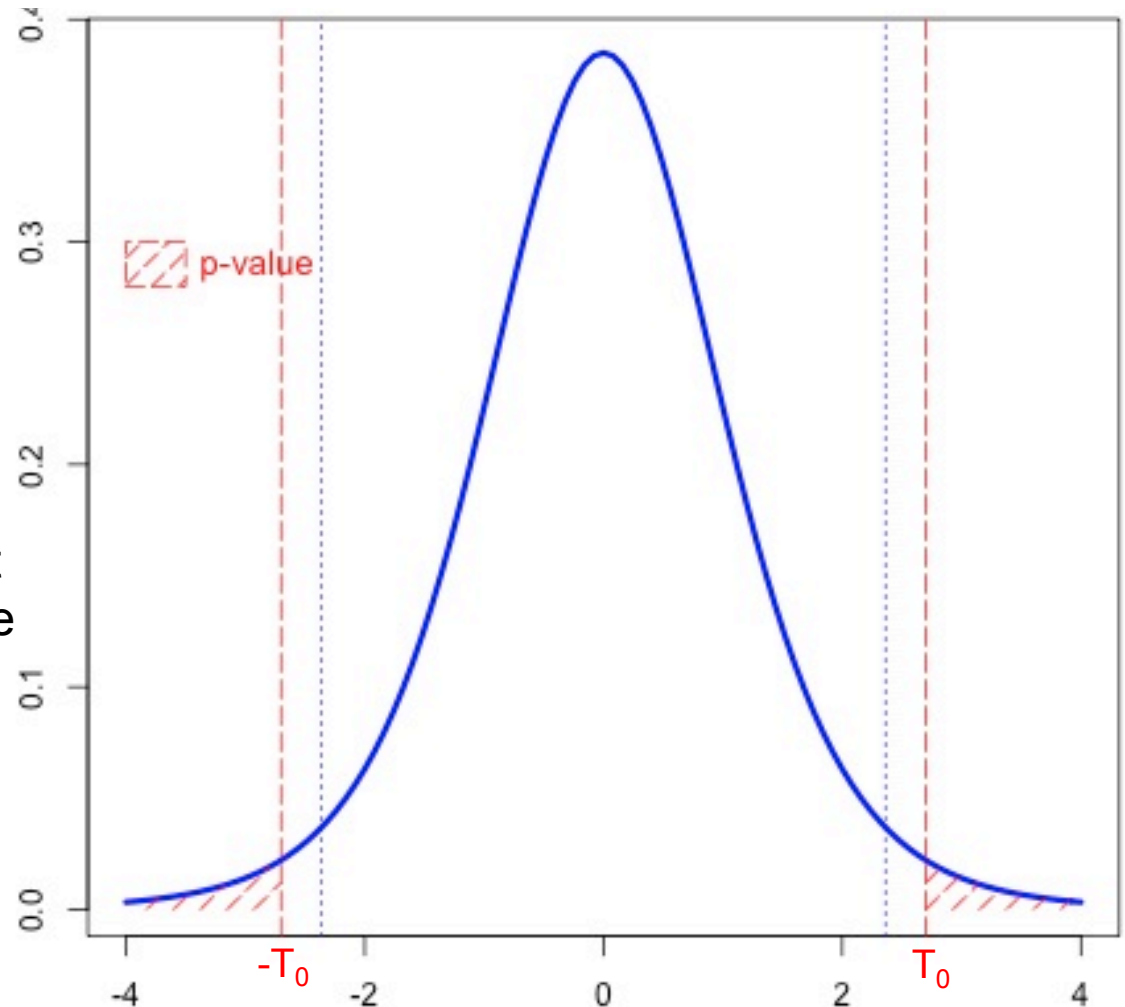
The P-Value:

Is the probability for T to have a value at least as **extreme** as T_0 by chance alone.

Tells how likely the value of T_0 is, assuming the null is true.

If it is very small, it would imply that the data obtained is less compatible with the null.

$$\begin{aligned} p\text{-value} &= P(T \leq -T_0 \text{ or } T > T_0) \\ &= P(|T| > T_0) \\ &= 2P(T \leq -T_0) \end{aligned}$$



Decision Rule, p-value approach

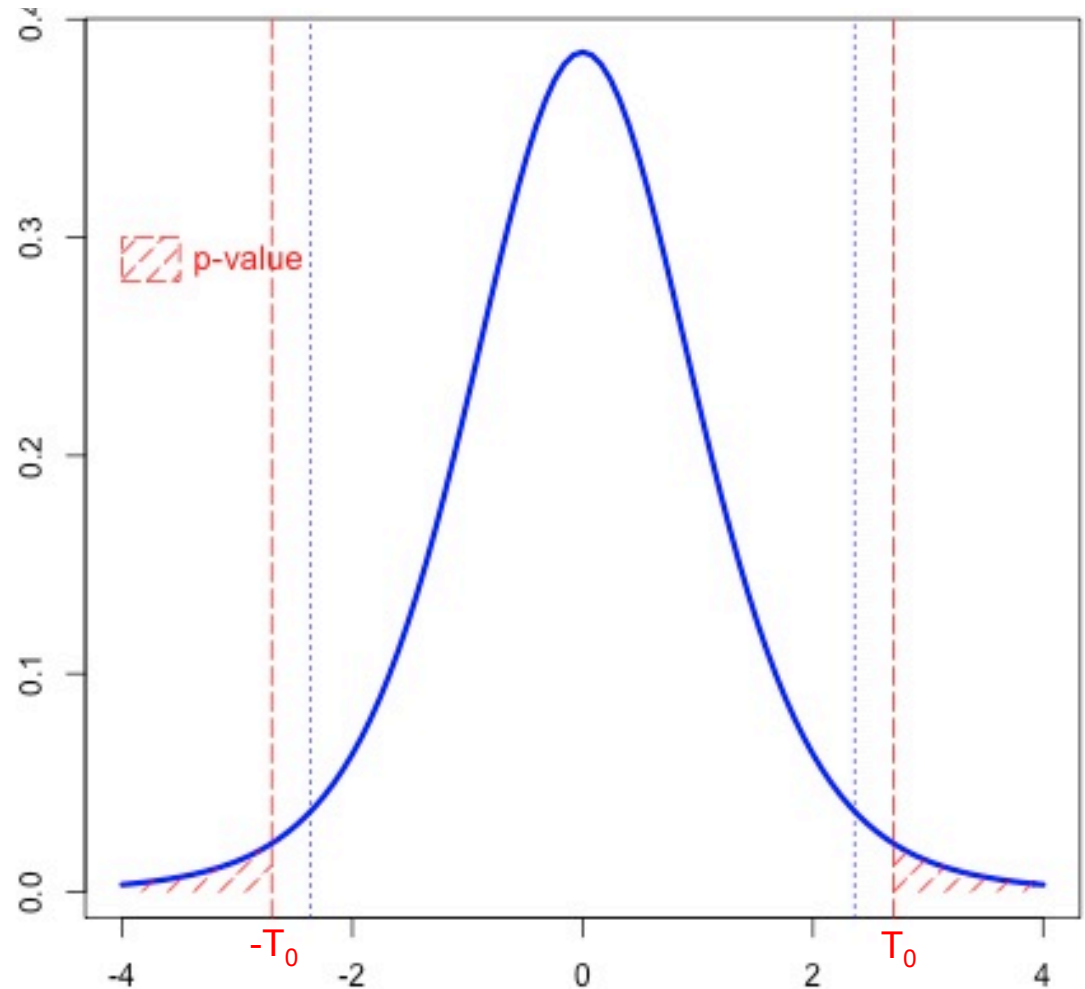
Two-tailed and One-tailed tests

Two-tailed test:

Extreme values of T can occur by chance at both tails of the distribution, which the p-value has to account for. This is the most common situation.

One –tailed test:

Where extreme values of T can only go in one direction. This is fairly uncommon if happened only by chance. P-values will be computed for one tail of the distribution.

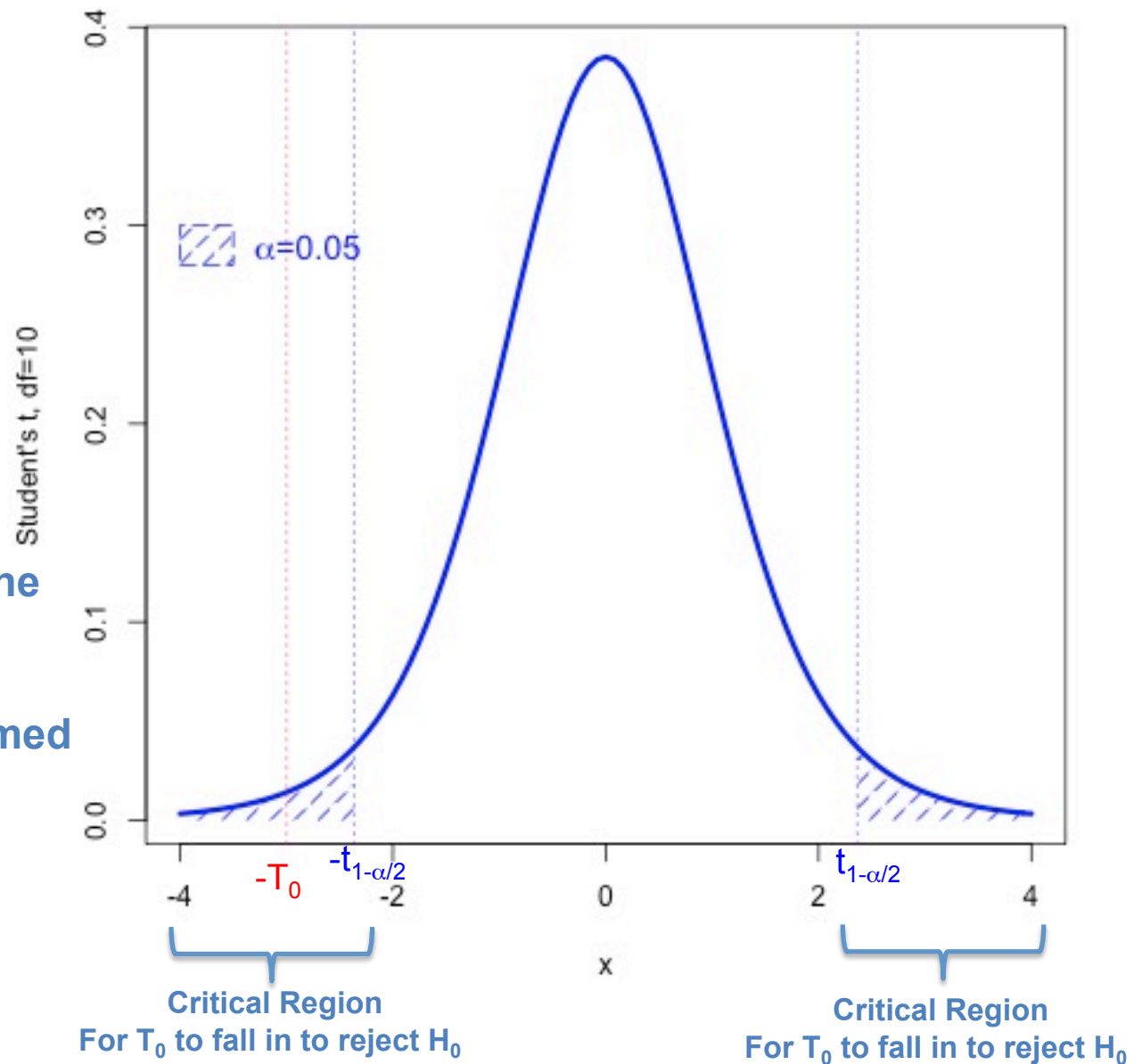


There is a second approach in assessing how extreme T_0 is.

Decision Rule, critical value/region approach

We reject H_0 if T_0 falls in the critical region.

At $\alpha=0.05$,
 $\pm t_{1-\alpha/2}$ are quantiles for the
2.5-th and 97.5-th
percentiles of the assumed
distribution under H_0 .



Reject H_0 at an α level if $T_0 \leq -t_{(1-\alpha)/2}$ or $T_0 > t_{(1-\alpha)/2}$... i.e., $|T_0| > t_{(1-\alpha)/2}$

Steps for Hypothesis Testing

1. Write null and alternative hypotheses statements.
2. Calculate the value of the appropriate test statistic
3. Assess how likely it is for the observed test statistic to occur. Apply decision rules via critical region and/or p-value.
4. Draw conclusions:
 - a) Is there enough evidence to reject H_0 ? At what α level?
 - b) What is the contextual significance?

Example: Plasma glucose level measurements (mmol/l) on 8 diabetic patients, before and one hour after administration of 100 g glucose (Altman, 1991).

One sample t-test

Decisions regarding the change in plasma glucose level after administration;

H_0 : there is no change in plasma glucose.

H_1 : there is change in plasma glucose.

Formally:

$$H_0 : \mu = 0 \quad \text{vs.} \quad H_1 : \mu \neq 0$$

where μ is the mean glucose change after treatment ($\mu_0 = 0$).

Patient	Plasma glucose (mmol/l)		
	Before	After	Change
1	4.67	5.44	0.77
2	4.97	10.11	5.14
3	5.11	8.49	3.38
4	5.17	6.61	1.44
5	5.33	10.67	5.34
6	6.22	5.67	-0.55
7	6.50	5.78	-0.72
8	7.00	9.89	2.89

Reminder

$$\text{Conceptual Test Statistic: } T = \frac{\text{statistic} - \text{postulated value}}{\text{SE}(\text{statistic})} \sim CDF$$

The conceptual test statistic

$$T = \frac{\bar{Y} - \mu_0}{s / \sqrt{n}}$$

The smaller the value of T, the more tenable H_0 is.

The assumed distribution of T, given the null is true:

$$T \sim \text{Student's } t \text{ (df} = n - 1\text{)}$$

df = n - #estimated parameters

The test statistic has a t-Distribution with n-1 degrees of freedom.

This distribution is used here because data is assumed to be approximately Normal with mean estimated by \bar{X} and σ unknown, estimated by s.

Reminder

$$\text{Observed Test Statistic: } T_0 = \frac{\text{Obs. statistic} - \text{postulated value}}{\text{obs. SE}(\text{statistic})}$$

Sample mean : $\bar{y} = 2.211$

Sample sd : $s = 2.363$

SE of sample mean, $n = 8$ $SE(\bar{y}) = s/\sqrt{n} = 0.835$

Observed test statistic

$$T_0 = \frac{\bar{y} - 0}{s/\sqrt{n}} = 2.647 \sim \text{Student's } t(df = 7)$$

We choose the α level to use when drawing conclusions. Here $\alpha=0.05$.

```
> t.test(dat$Change)
```

One Sample t-test

data: dat\$Change

t = 2.6469, df = 7, p-value = 0.03309

alternative hypothesis: true mean is not equal to 0

95 percent confidence interval:

0.2358411 4.1866589

sample estimates:

mean of x

2.21125

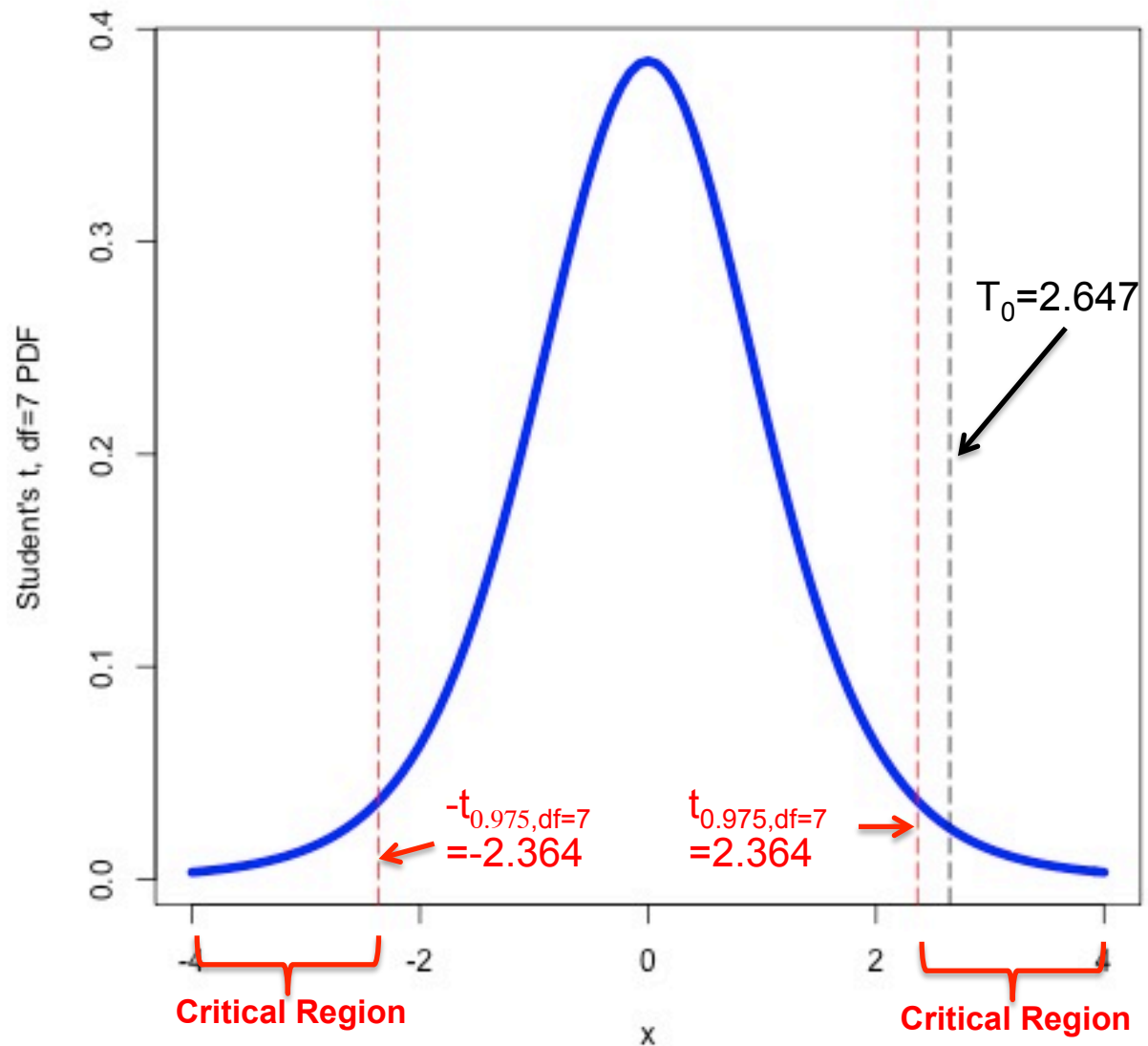
```
> 2*pt(-2.6469,df=7)
[1] 0.03309027
> 2*(1-pt(2.6469,df=7))
[1] 0.03309027
```

“There is a statistically significant change in mean plasma glucose after one hour of glucose administration (mean change=2.21, p-val=0.03)”.

Since $\alpha=0.05$, The critical region can be determined by the quantiles for the 2.5-th and 97.5-th percentiles of the t-Distribution with $df=n-1$.

```
> qt(.975,df=7)  
[1] 2.364624
```

Does $T_0=2.647$,
fall in the critical
region?



Example: Body composition measurements (percent fat) on 25 normal adults, men and women, between 23 and 61 years old. (Altman, 1991).

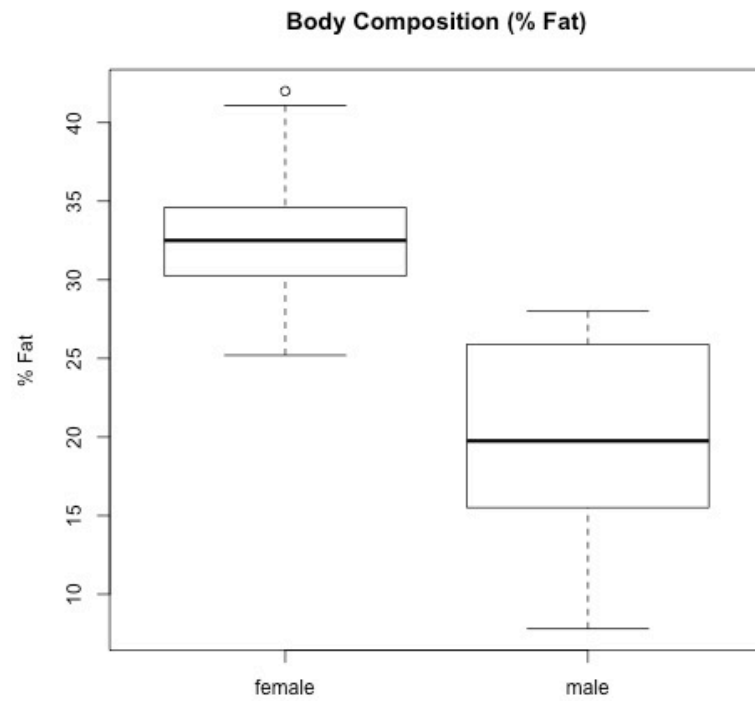
“Two sample t-test”

Data is built in the HSAUR R package.

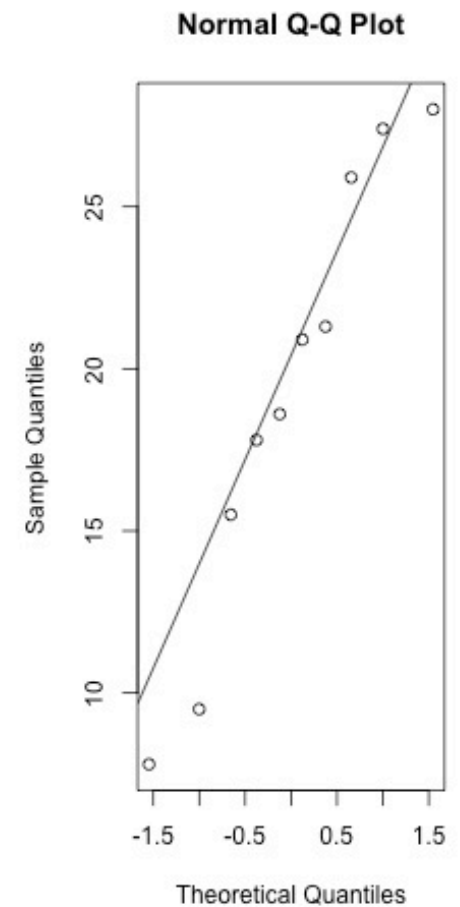
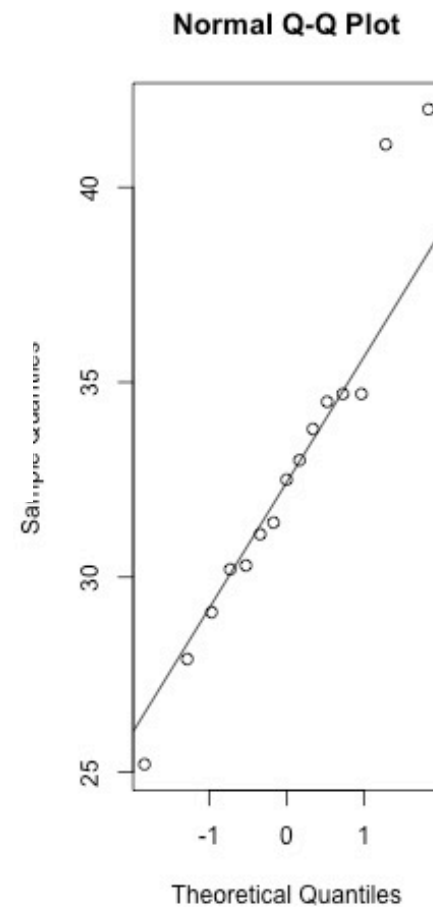
```
> library(HSAUR)
> head(agefat)
  age  fat  sex
1  24 15.5 male
2  37 20.9 male
3  41 18.6 male
4  60 28.0 male
5  31 34.7 female
6  39 30.2 female
```

```
> str(agefat)
'data.frame':  25 obs. of  3 variables:
 $ age: int  24 37 41 60 31 39 58 23 23 27 ...
 $ fat: num  15.5 20.9 18.6 28 34.7 30.2 21.3 9.5 27.9 7.8 ...
 $ sex: Factor w/ 2 levels "female","male": 2 2 2 2 1 1 2 2 1 2 ...
```

Exploratory view of Body Composition (% Fat) and Normality assessment



Sample size:
female male
15 10



Reminder

Conceptual Test Statistic: $T = \frac{\text{statistic} - \text{postulated value}}{\text{SE}(\text{statistic})} \sim CDF$

$$H_0: \mu_{\text{female}} = \mu_{\text{male}} \quad \text{vs.} \quad H_1: \mu_{\text{female}} \neq \mu_{\text{male}}$$

$$H_0: \mu_{\text{female}} - \mu_{\text{male}} = 0 \quad \text{vs.} \quad H_1: \mu_{\text{female}} - \mu_{\text{male}} \neq 0$$

Test statistic for the mean difference in %Fat

With pooled variance

$$T = \frac{\bar{Y}_1 - \bar{Y}_2}{\sqrt{s_P^2 \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}} \sim \text{Student's } t(n-2)$$

$$s_P^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$

The pooled variance formula is based on the average of the two variances, giving more weight to the larger sample. Here we assume that

- (i) the population means are different but the variances are the same, given by S_p^2
- (ii) the two samples are independent.

$$\text{Var}(\bar{Y}_1 - \bar{Y}_2) = \text{Var}(\bar{Y}_1) + \text{Var}(\bar{Y}_2) = \frac{s_P^2}{n_1} + \frac{s_P^2}{n_2} = s_P^2 \left(\frac{1}{n_1} + \frac{1}{n_2} \right).$$

Reminder

$$\text{Observed Test Statistic: } T_0 = \frac{\text{Obs. statistic} - \text{postulated value}}{\text{obs. SE}(\text{statistic})}$$

```
> t.test(fat~sex,var.equal=T,data=agefat)
```

Two Sample t-test

```
data: fat by sex
t = 5.9296, df = 23, p-value = 4.803e-06
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 8.788105 18.205228
sample estimates:
mean in group female    mean in group male
      32.76667           19.27000
```

“The difference in mean body fat percentage between men and women is highly statistically significant (mean difference=13.5%, p-val=4.8e-6)”.

We choose $\alpha=0.05$

R Code
Two sample test for the difference of means
(manual computation)

```
library(HSAUR)

agefat

attach(agefat)

means <- tapply(fat,sex,mean)

sds <- tapply(fat,sex,sd)

n <- table(sex)

meandiff <- as.numeric(means[1] - means[2])

pooledS2.numerator <- (n[1]-1)*sds[1]^2 + (n[2]-1)*sds[2]^2
pooledS2.denominator <- n[1]+n[2]-2
pooledS2 <- as.numeric( pooledS2.numerator / pooledS2.denominator )

SE.meandiff <- as.numeric( sqrt(pooledS2*(1/n[1]+1/n[2])) )

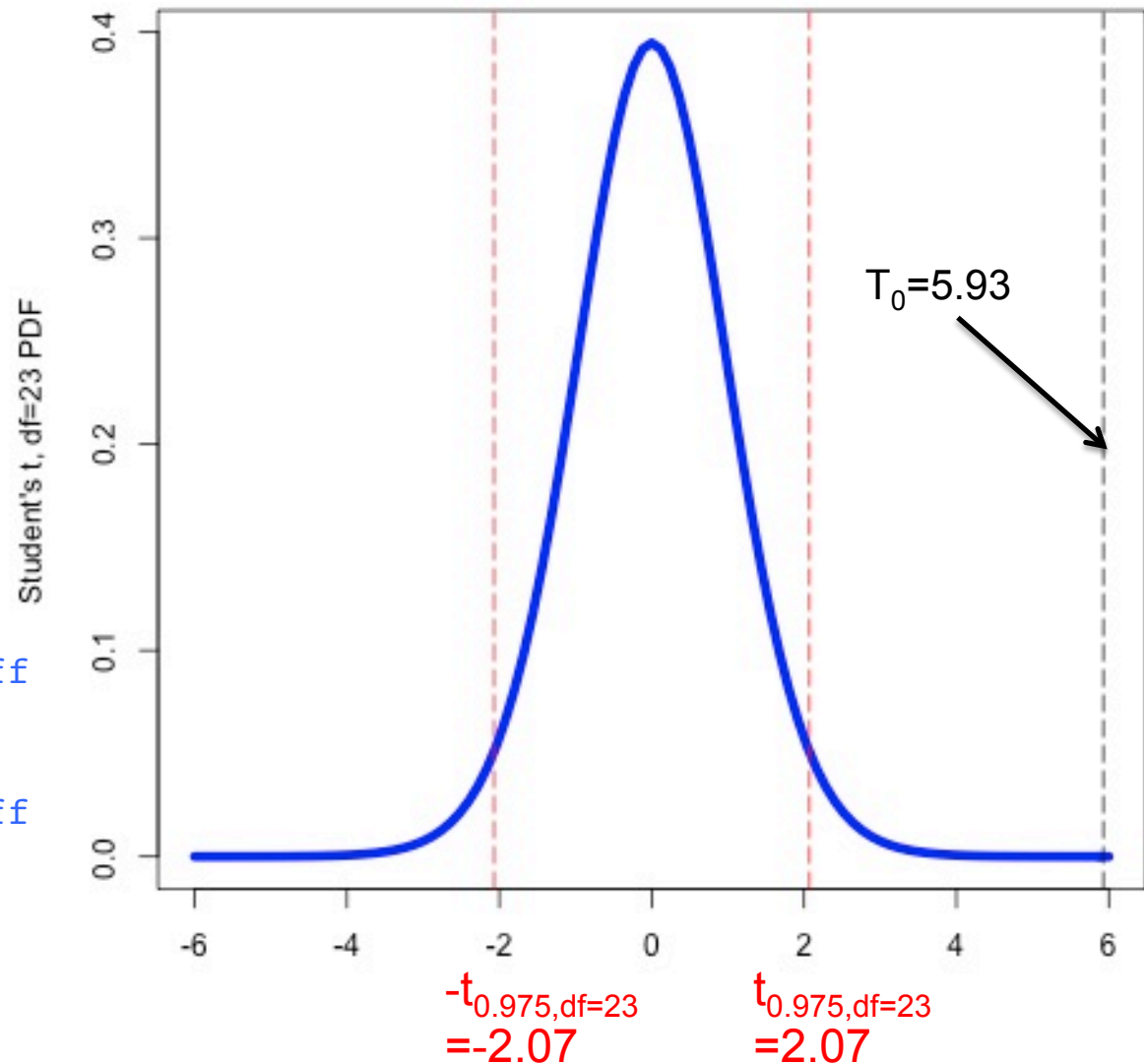
t.statistic <- meandiff/SE.meandiff
```

R output

Two sample test for the difference of means

```
> means
  female    male
32.76667 19.27000
> meandiff
[1] 13.49667
> t.statistic
[1] 5.929621

> # p-value:
> 2*pt(-t.statistic,df=23)
[1] 4.803389e-06
> # 95% CI
> meandiff-qt(.975,df=23)
               *SE.meandiff
[1] 8.788105
> meandiff+qt(.975,df=23)
               *SE.meandiff
[1] 18.20523
> qt(.975,df=23)
[1] 2.068658
```



Summary: Generic components of hypothesis testing

H_0 : parameter = postulated value

vs.

H_1 : parameter \neq postulated value \leftarrow Two tailed

or H_1 : parameter $> (<)$ postulated value \leftarrow One tailed

Hypothesis
statement

Test
statistic

Conceptual,
theoretical test
statistic

$$T = \frac{(\text{statistic} - \text{postulated value})}{\text{SE}(\text{statistic})}$$

Observed,
sample-based
test statistic

$$T_0 = \frac{(\text{observed statistic} - \text{postulated value})}{\text{observed SE}(\text{statistic})}$$

Sampling
distribution of T

$$T \sim CDF$$

Decision rule
at an α level

Reject H_0 if

$P(T_0 \text{ being an extreme value}) \leq \alpha$

P-value approach

T_0 is inside the critical
region given by $\pm t_{1-\alpha/2}$

Critical region approach

Final note on CI's interpretation

Picky, picky, picky! A 95% chance of what?

It is correct to say that there is a 95% chance that the **CI contains** the true population mean.

It is not quite correct to say that there is a 95% chance that **the population mean lies** within the interval.

What's the difference?

The population mean has one value. You don't know what it is. If you repeated the experiment, that value wouldn't change.

Therefore it isn't strictly correct to ask about the probability that the population mean lies within a certain range. The population mean is not a random variable.

In contrast, the confidence interval you compute depends on the data you happened to collect. The CI does randomly vary?