Review of Basic Statistical Concepts
M.Sc. Politics and Policy Analysis

Francesco Mattioli

francesco.mattioli@unibocconi.it

Outline

Why do we need to talk about *probability*?

- Most aspects of the world around us have an element of randomness
- **Probability theory**: Quantifying Randomness
- Some Definitions:
  - **Outcome (y)**: (mutually exclusive) result of a Random Process
  - **Probability (p)**: proportion of times that a certain outcome is observed if you repeat a random process many times
  - **Random Variable (Y)**: variable (discrete or continuous) that can take on a set of different values, each with an associated probability

Probability Distributions – Discrete Random Variables

Probability Distribution or **Probability Mass Function**: list of all possible values y_i, with $i = 1, ..., N$, of the random variable Y and the probability that each value occurs (frequency): $f(Y) = Pr(Y = y_i)$

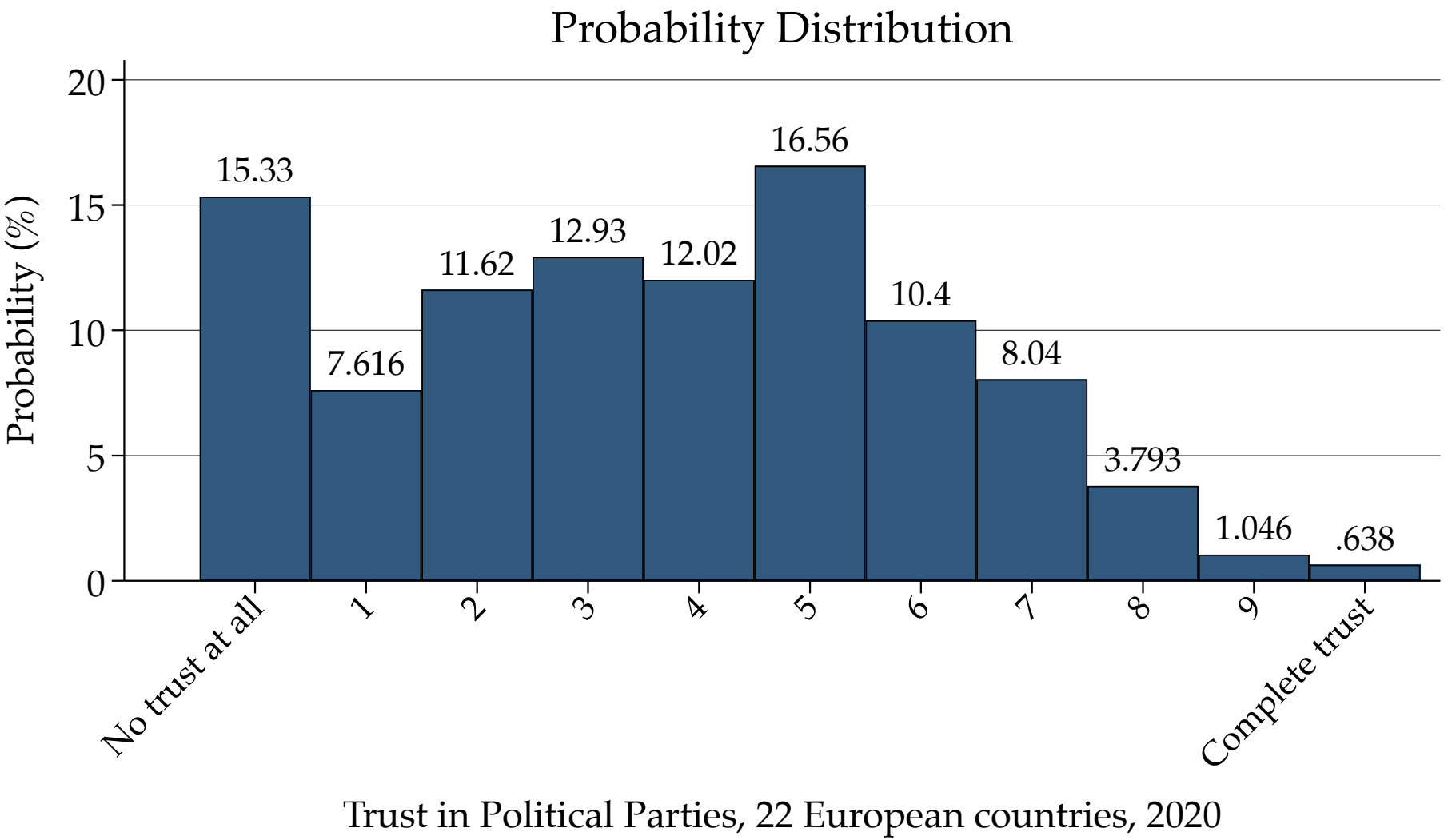
Example. European Social Survey, Wave 10 (free to download upon registration)

```
. tabulate trstprt
```

Y Trust in political parties	Freq.	Percent	Cum.
No trust at all	5,671	15.33	15.33
1	2,817	7.62	22.95
2	4,299	11.62	34.57
3	4,784	12.93	47.50
4	4,445	12.02	59.52
5	6,127	16.56	76.09
6	3,845	10.40	86.48
7	2,974	8.04	94.52
8	1,403	3.79	98.32
9	387	1.05	99.36
Complete trust	236	0.64	100.00
Total	36,988	100.00	

(y_i = No trust at all … Complete trust; $f(Y)$ = Percent)

Probability Distributions – Graphical Representation



Trust in Political Parties, 22 European countries, 2020

Probability Distributions – Discrete Random Variables

Cumulative Distribution Function (cdf): probability that the random variable is less than or equal to a given value:

$$F(Y) = Pr(Y \leq y_k) = \sum_{i=1}^{k} f(y_i)$$

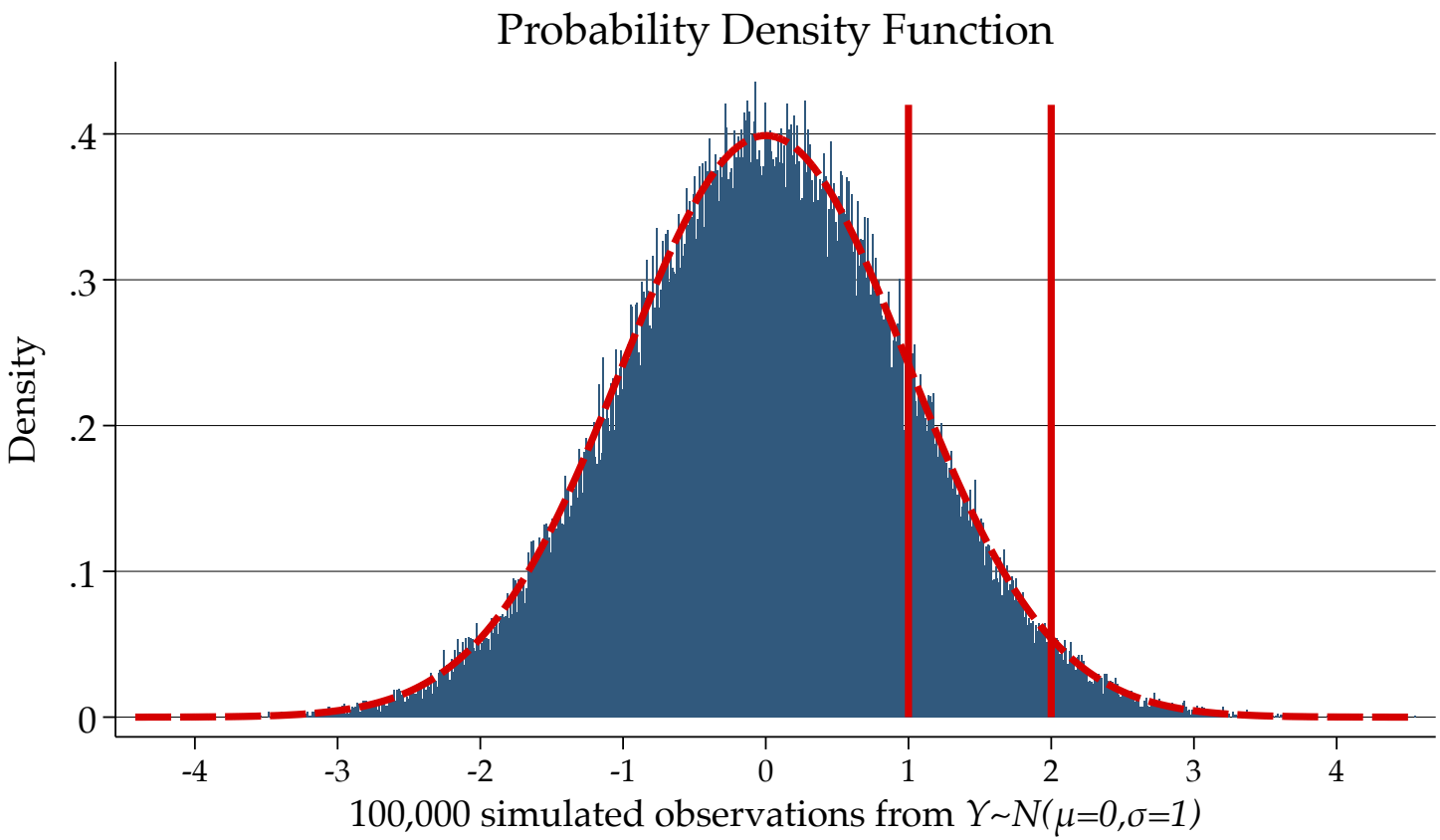
```
. tabulate trstprt
```

Y Trust in political parties	Freq.	Percent $f(Y)$	Cum. $F(Y)$
No trust at all	5,671	15.33	15.33
1	2,817	7.62	22.95
2	4,299	11.62	34.57
3	4,784	12.93	47.50
4	4,445	12.02	59.52
5	6,127	16.56	76.09
6	3,845	10.40	86.48
7	2,974	8.04	94.52
8	1,403	3.79	98.32
9	387	1.05	99.36
Complete trust	236	0.64	100.00
Total	36,988	100.00	

(y_i)

Histograms – Continuous Random Variables

Probability Density Function (pdf): area under pdf between two values is the probability that the random variable falls between those two values: $f(Y)$



Moments of a distribution

Measures of Central Tendency

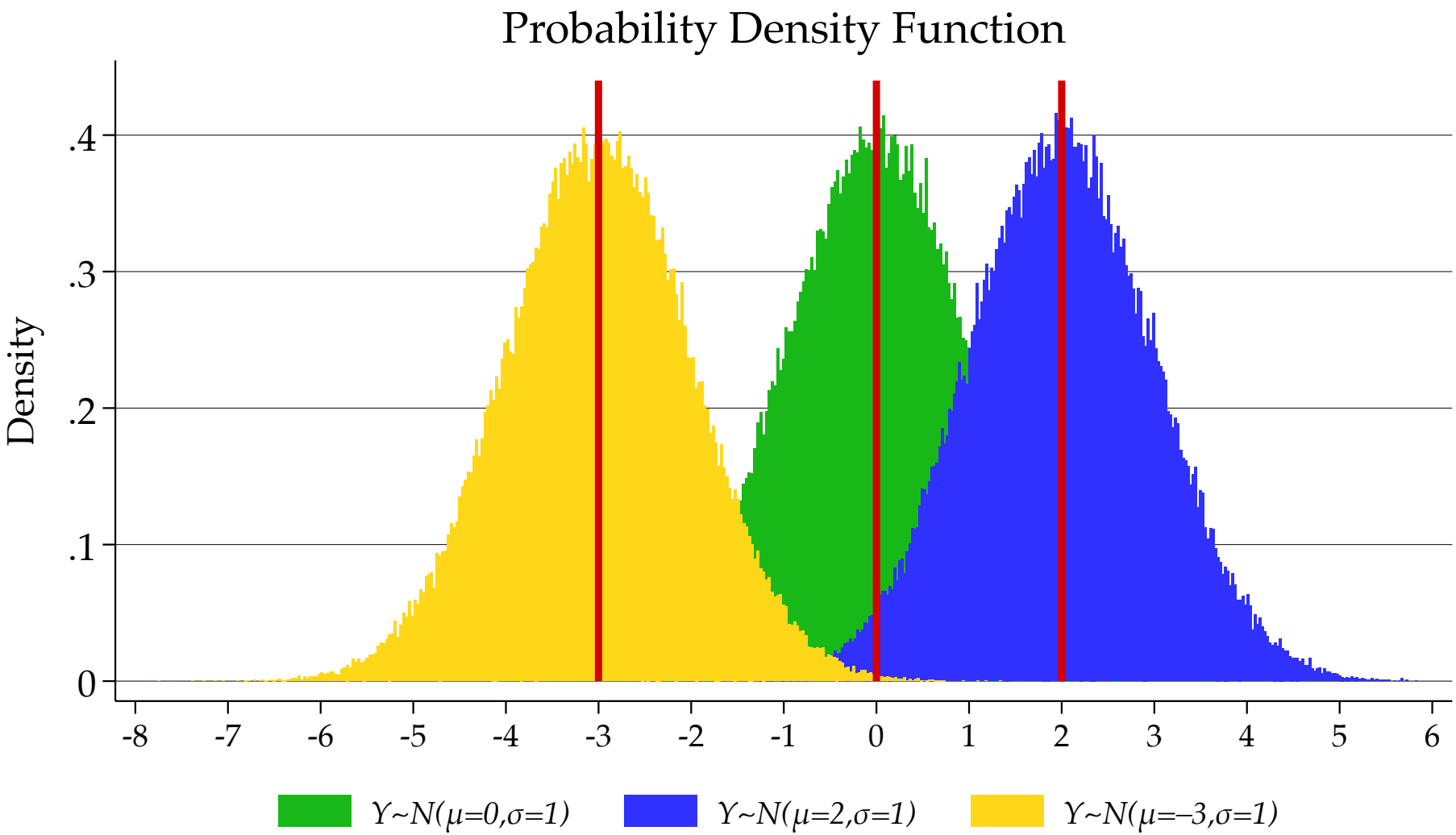
- **Expected Value**, $E(Y)$, or **Mean**, μ_Y (population mean) or $\overline{Y}$ (sample mean) – First moment
- **Median**: the value y_m that splits the distribution in two equal parts (50% of the distribution on its left, 50% on its right)
- **Mode**: the value with the highest frequency

Example: Mean computation

$$\overline{Y} = p_1 y_1 + p_2 y_2 + ... + p_N y_N = \sum_{i=1}^{N} p_i y_i$$

y_i	p_i	$p_i y_i$
0	0.1533	0
1	0.0762	0.0762
2	0.1162	0.2325
3	0.1293	0.3880
4	0.1202	0.4807
5	0.1656	0.8282
6	0.1040	0.6237
7	0.0804	0.5628
8	0.0379	0.3034
9	0.0105	0.0942
10	0.0064	0.0638
		$\overline{Y}$ 3.6535

Comparing distributions with different means



Moments of a distribution

Measures of Dispersion (second moment)

- (Population) **Variance**

$$Var(Y) = \sigma_Y^2 = \sum_{i=1}^{N} p_i (y_i - \overline{Y})^2$$

$$= p_1 (y_1 - \overline{Y})^2 + p_2 (y_2 - \overline{Y})^2 + ... + p_N (y_N - \overline{Y})^2$$

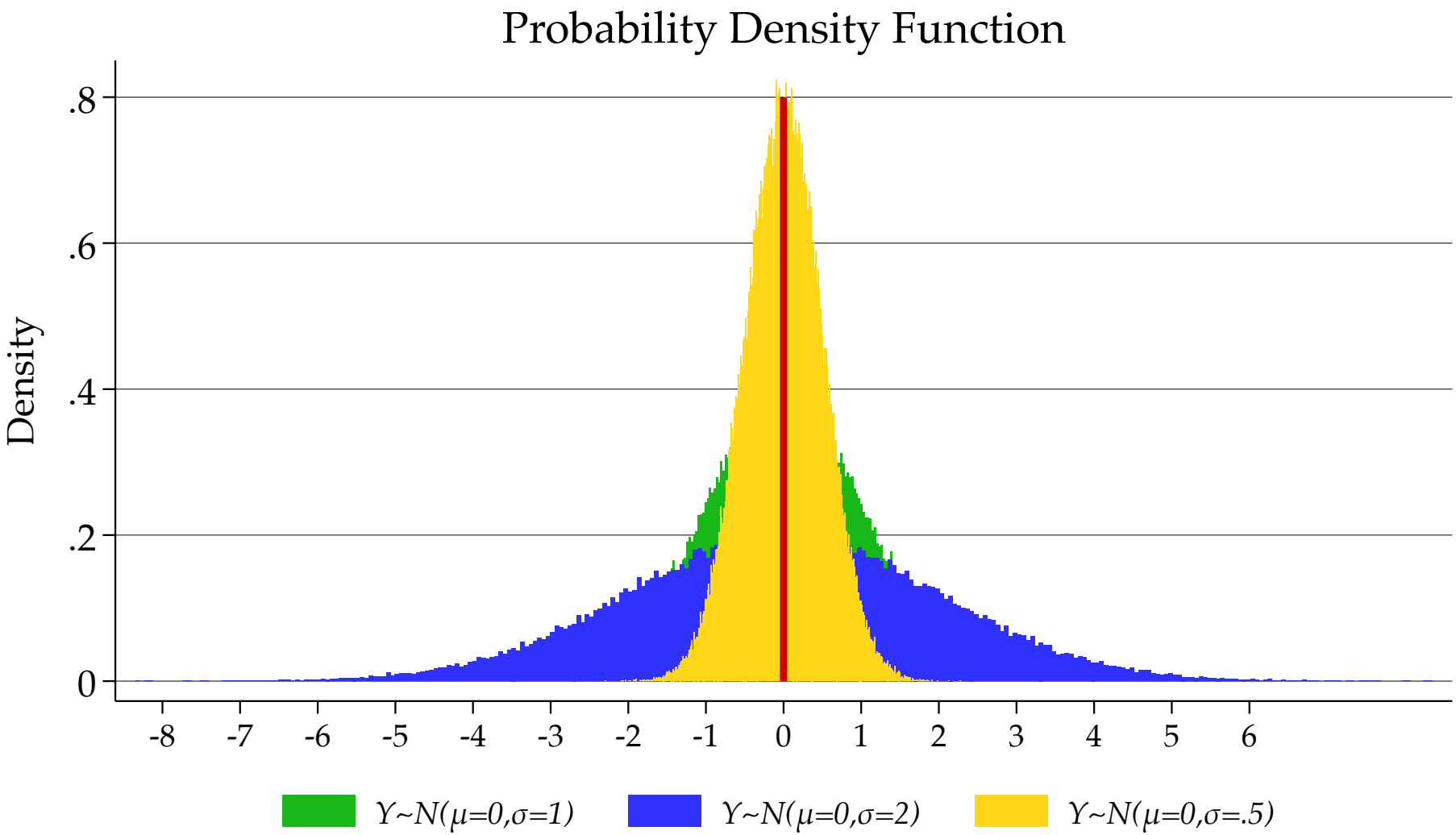
- **Standard Deviation**

$$\sigma_Y = \sqrt{Var(Y)}$$

Example: Variance computation

y_i	$\overline{Y}$	$(y_i - \overline{Y})^2$	p_i	$p_i(y_i - \overline{Y})^2$
0	3.6535	13.3483	0.1533	2.0466
1	3.6535	7.0413	0.0762	0.5363
2	3.6535	2.7342	0.1162	0.3178
3	3.6535	0.4271	0.1293	0.0552
4	3.6535	0.1200	0.1202	0.0144
5	3.6535	1.8130	0.1656	0.3003
6	3.6535	5.5059	0.1040	0.5724
7	3.6535	11.1988	0.0804	0.9004
8	3.6535	18.8917	0.0379	0.7166
9	3.6535	28.5847	0.0105	0.2991
10	3.6535	40.2776	0.0064	0.2570
			σ_Y^2	6.0160
			σ_Y	2.4528

Comparing distributions with different variance



Quantiles

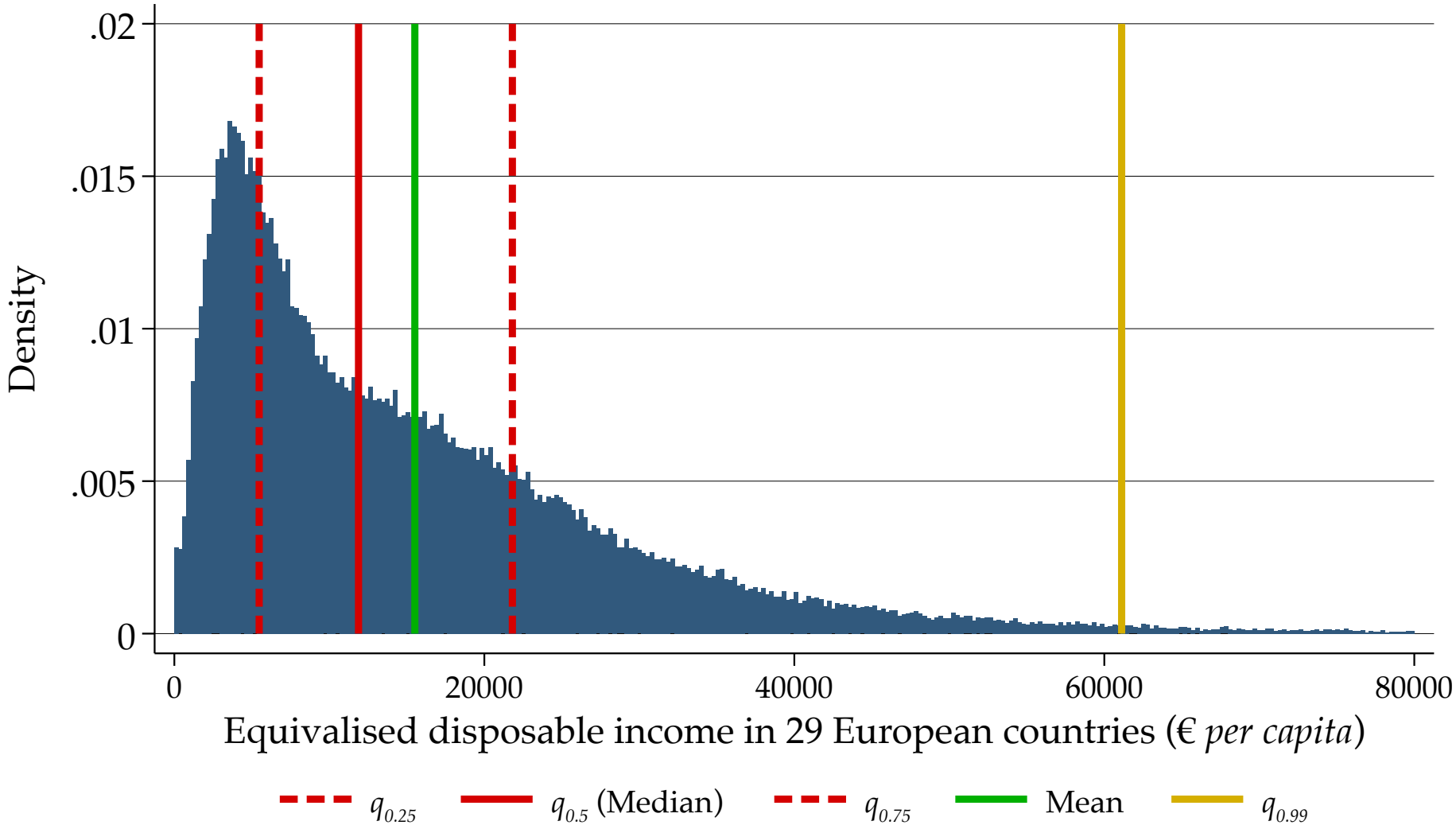
$$q_\phi = min\{y : F(y) \geq \phi\}$$

Very important quantiles:

- Median: $q_{0.5}$
- Tertiles: $q_{0.33}, q_{0.66}$
- Quartiles: $q_{0.25}, q_{0.5}, q_{0.75}$
- Quintiles: $q_{0.2}, q_{0.4}, q_{0.6}, q_{0.8}$
- Deciles: $q_{0.1}, q_{0.2}, ..., q_{0.9}$
- Percentiles: $q_{0.01}, q_{0.02}, ..., q_{0.99}$

Example: Quantiles in the income distribution

EU Survey on Income and Living Conditions (EU-SILC)



Example: Quantiles in the income distribution

EU Survey on Income and Living Conditions (EU-SILC)

```
. summarize hy020_pc, detail
```

hy020_pc

Percentiles		Smallest		
1%	**829.7266**	**0**		
5%	**1999.6**	**0**		
10%	**2949.156**	**0**	Obs	**150,173**
25%	**5471.836**	**0**	Sum of wgt.	**150,173**
50%	**11885**		Mean	**15515.85**
		Largest	Std. dev.	**13116.37**
75%	**21810**	**79829**		
90%	**33109.45**	**79883.87**	Variance	**1.72e+08**
95%	**41715**	**79924.28**	Skewness	**1.497488**
99%	**61120**	**79996.52**	Kurtosis	**5.622403**

Concepts involving two random variables

Joint Probability Distribution: the joint probability distribution of two random variables X and Y is the probability that the Y and X simultaneously take on certain values y_i and x_j:
$f(y_i, x_j) = Pr(Y = y_i, X = x_j)$

Example: joint distribution of Y (Trust in political parties – recoded) and X (Voted in last election)

```
. tabulate trstprt_3 vote, cell
```

trstprt_3	Voted last national election: Yes	No	Not eligi	Total
Low	11,886	4,631	805	17,322
	32.50	12.66	2.20	47.36
Medium	10,836	2,342	1,099	14,277
	29.63	6.40	3.01	39.04
High	3,839	630	504	4,973
	10.50	1.72	1.38	13.60
Total	26,561	7,603	2,408	36,572
	72.63	20.79	6.58	100.00

Concepts involving two random variables

Conditional Distribution: the conditional distribution of one variable Y given another variable X is the distribution of Y conditional on X taking on a specific value x_k: $f(Y|X) = Pr(Y = y_i|X = x_k)$

Example: conditional distribution of Y (Trust in political parties – recoded) given certain values of X (Voted in last election)

```
. tabulate trstprt_3 if vote==1
```

trstprt_3	Freq.	Percent	Cum.
Low	**11,886**	**44.75**	**44.75**
Medium	**10,836**	**40.80**	**85.55**
High	**3,839**	**14.45**	**100.00**
Total	**26,561**	**100.00**	

Concepts involving two random variables

Conditional Distribution: the conditional distribution of one variable Y given another variable X is the distribution of Y conditional on X taking on a specific value x_k: $f(Y|X) = Pr(Y = y_i|X = x_k)$

Example: conditional distribution of Y (Trust in political parties – recoded) given certain values of X (Voted in last election)

```
. tabulate trstprt_3 if vote==2
```

trstprt_3	Freq.	Percent	Cum.
Low	4,631	60.91	60.91
Medium	2,342	30.80	91.71
High	630	8.29	100.00
Total	7,603	100.00	

Concepts involving two random variables

Conditional Mean: the mean of Y conditional on X taking on a specific value x_k:

$$\overline{Y} = f(y_1|X = x_k) \cdot y_1 + ... + f(y_N|X = x_k) \cdot y_N = \sum_{i=1}^{N} f(y_i|X = x_k) \cdot y_i$$

Example: conditional mean of Y (Trust in political parties – recoded) given certain values of X (Voted in last election)

```
. bysort vote: summarize trstprt_3

-------------------------------------------------------------------------------
-> vote = Yes

    Variable |        Obs        Mean    Std. dev.       Min        Max
-------------+---------------------------------------------------------
   trstprt_3 |     26,561    1.697037    .7072946         1          3

-------------------------------------------------------------------------------
-> vote = No

    Variable |        Obs        Mean    Std. dev.       Min        Max
-------------+---------------------------------------------------------
   trstprt_3 |      7,603     1.47376    .6442749         1          3

-------------------------------------------------------------------------------
-> vote = Not

    Variable |        Obs        Mean    Std. dev.       Min        Max
-------------+---------------------------------------------------------
   trstprt_3 |      2,408       1.875     .726773         1          3
```

Concepts involving two random variables

Independence: two random variables X and Y are independently distributed if knowing the value of X provides no information about Y: $f(Y|X) = Pr(Y = y_i|X = x_k) = Pr(Y = y_i)$

Example: consider again Y and X; are they independent? **No**. (i) joint prob. $\neq$ product of marginal prob. (ii) conditional prob. $\neq$ marginal prob.

```
. tabulate trstprt_3 vote, cell column
```

trstprt_3	Voted last national election: Yes	No	Not eligi	Total
Low	11,886	4,631	805	17,322
	44.75	60.91	33.43	47.36
	32.50	12.66	2.20	47.36
Medium	10,836	2,342	1,099	14,277
	40.80	30.80	45.64	39.04
	29.63	6.40	3.01	39.04
High	3,839	630	504	4,973
	14.45	8.29	20.93	13.60
	10.50	1.72	1.38	13.60
Total	26,561	7,603	2,408	36,572
	100.00	100.00	100.00	100.00
	72.63	20.79	6.58	100.00

Concepts involving two random variables

Covariance: measure of the extent to which two random variables move together:

$$Cov(Y, X) = \sigma_{YX} = \sum_{i=1}^{N}\sum_{j=1}^{K}(y_i - \overline{Y}) \cdot (x_j - \overline{X}) \cdot f(Y = y_i, X = x_j)$$

Example: covariance between Y and X

Variable	Obs	Mean	Std. dev.	Min	Max
trstprt_3	36,572	1.662337	.7040006	1	3
vote	36,572	1.339577	.5966235	1	3

```
. tabulate trstprt_3 vote, cell
```

trstprt_3	Voted last national election: Yes	No	Not eligi	Total
Low	11,886	4,631	805	17,322
	32.50	12.66	2.20	47.36
Medium	10,836	2,342	1,099	14,277
	29.63	6.40	3.01	39.04
High	3,839	630	504	4,973
	10.50	1.72	1.38	13.60
Total	26,561	7,603	2,408	36,572
	72.63	20.79	6.58	100.00

Concepts involving two random variables

Example: covariance between Y and X

$$Cov(Y,X) = \sigma_{YX} = \sum_{i=1}^{N}\sum_{j=1}^{K}(y_i - \overline{Y}) \cdot (x_i - \overline{X}) \cdot f(Y = y_i, X = x_i)$$

$$Cov(Y,Y) = \sigma_Y^2 = \sum_{i=1}^{N}(y_i - \overline{Y}) \cdot f(Y = y_i)$$

$$Cov(X,X) = \sigma_X^2 = \sum_{j=1}^{K}(x_j - \overline{X}) \cdot f(X = x_j)$$

```
. correlate trstprt_3 vote, covariance
(obs=36,572)
```

	trstpr~3	vote
trstprt_3	**.495617**	
vote	**-.011199**	**.35596**

Concepts involving two random variables

Correlation: (standardized) measure of the extent to which two random variables move together:

$$Corr(Y, X) = \frac{\sigma_{YX}}{\sigma_Y \cdot \sigma_X} = \frac{Cov(Y, X)}{\sqrt{Var(Y) \cdot Var(X)}}$$

- $-1 \leq Corr(Y, X) \leq 1$
- $Corr(Y, X) = 0 \rightarrow X$ and Y are uncorrelated

Example: correlation between Y and X

```
. correlate trstprt_3 vote
(obs=36,572)
```

	trstpr~3	vote
trstprt_3	**1.0000**	
vote	**-0.0267**	**1.0000**

Why do we need *Statistics*?

We cannot run a survey of a full population whenever we want to answer questions about unknown characteristics of its distribution.

Statistical Inference: we can learn about a characteristics of a population by selecting a random sample of that population

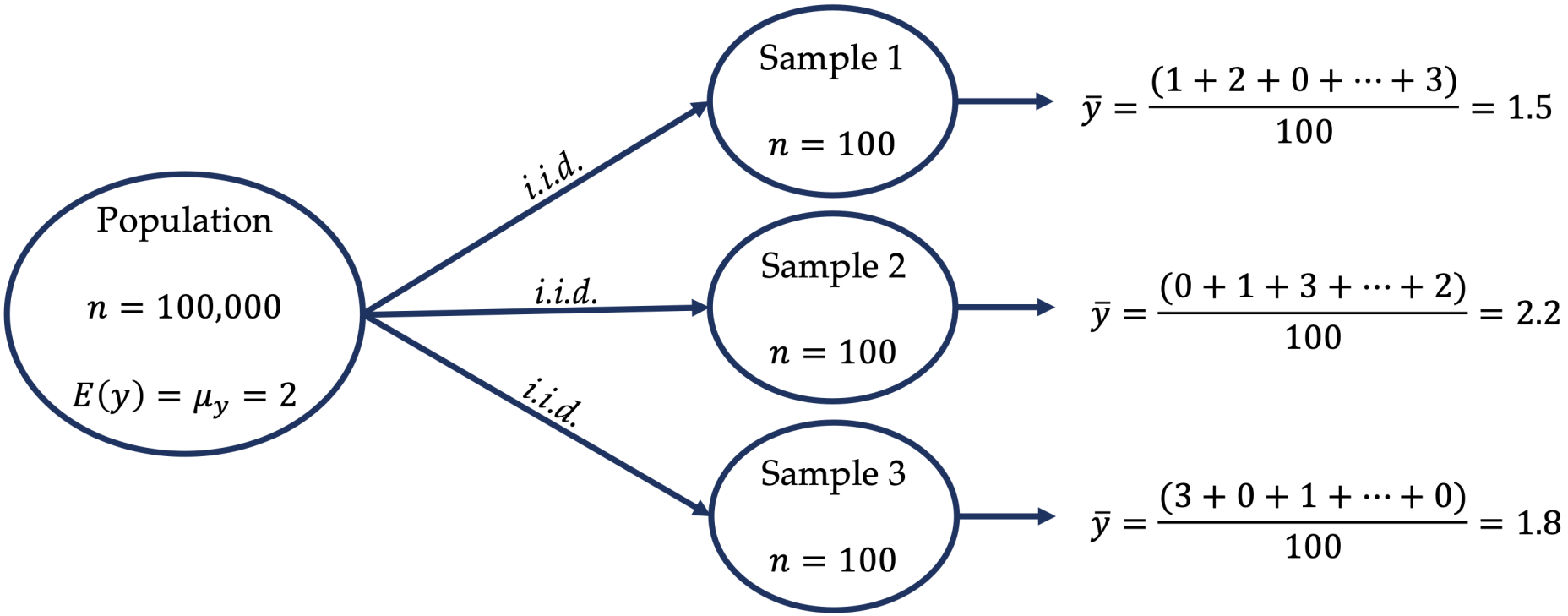
Econometrics uses three main statistical methods:

- **Estimation**: computing a *Best Guess* numerical value for an unknown characteristic (*parameter*) of a population distribution, from a sample of data
- **Hypothesis testing**: formulating a hypothesis about the population and use sample evidence to decide if it is true
- **Confidence Intervals**: use the sample data to calculate a range of statistically plausible values around the best guess for the unknown population characteristic

Estimators and their Properties

- We want to know the mean value of y in a population (μ_y is the parameter to be estimated)
- Draw a random sample of n independently and identically distributed (*iid*) observations $y_1, y_2, ..., y_n$
- Compute the sample average $\overline{y} = \frac{y_1 + ... + y_n}{n}$
- $\overline{y}$ is an **estimator** of μ_y (a function of the sample)
- $\overline{y}$ is a random variable, because it is influenced by the random draw of the sample (the individual you draw as first or i^{th} observation y_i is random!)
- The **estimate** (the actual value that $\overline{y}$ takes) is not random variable, but a scalar (a number)
- If you repeat the random draw from the same population a second time, the same **estimator** (random variable) $\overline{y}$ will produce a different **estimate** (scalar)
- As all random variables $\overline{y}$ has a probability distribution called **sampling distribution**

Example. Sampling distribution of $\bar{y}$

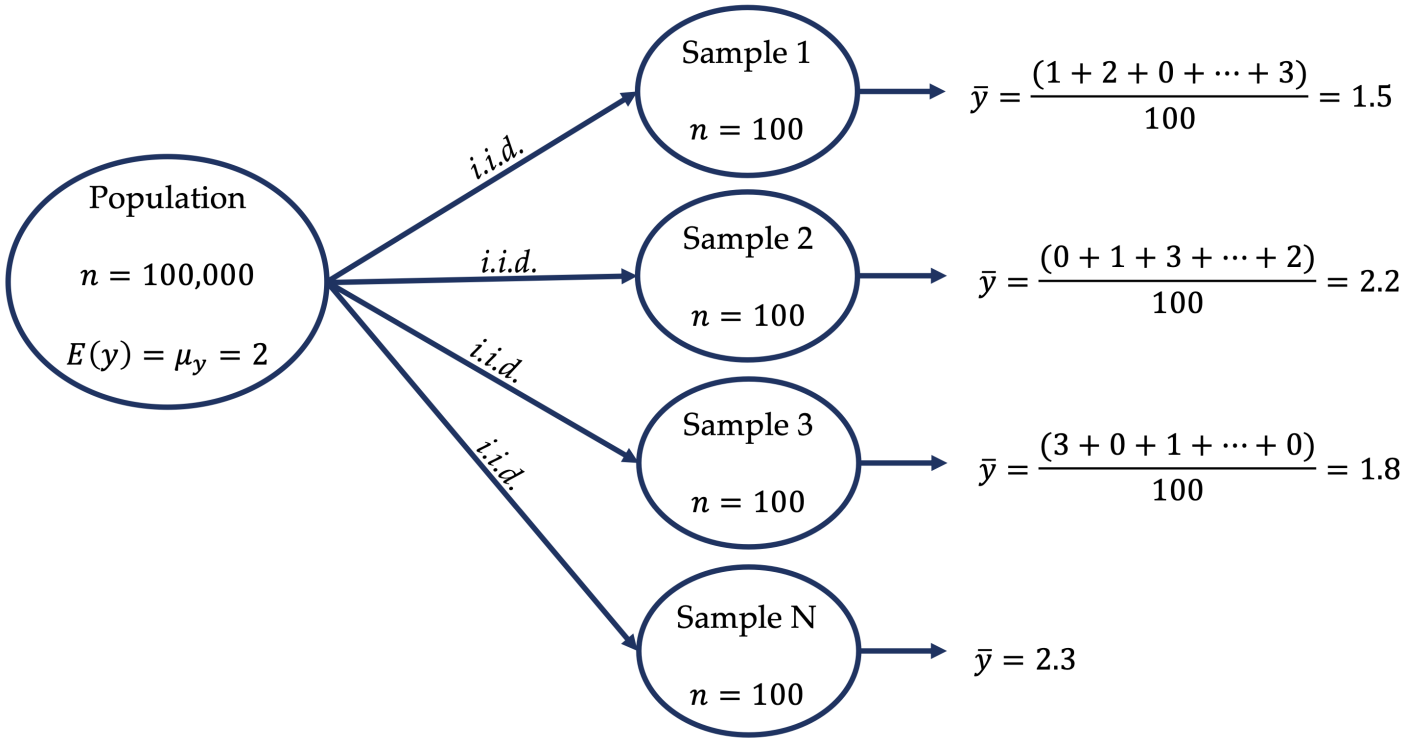


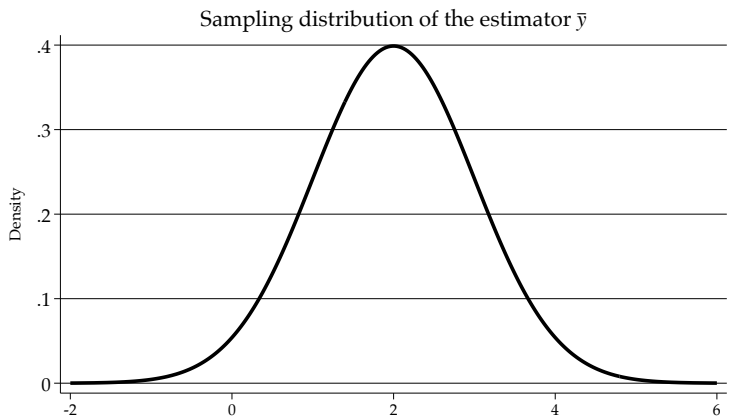
$\bar{y}$	$f(\bar{y})$
1.5	1/3
2.2	1/3
1.8	1/3

$$E(\bar{y}) = 1.5 \cdot 1/3 + 2.2 \cdot 1/3 + 1.8 \cdot 1/3 = 1.833$$

Example. Sampling distribution of $\bar{y}$

Estimator (random variable) $\bar{y}$ has its sampling distribution





Estimators and their Properties

What makes an estimator 'good'? A good estimator gets as close as possible to the unknown true value of the population parameter. Desirable properties of a good estimator:

- **Unbiasedness**: an estimator is **biased** if it's different, on average, from the true value of the parameter that is being estimated; draws from the same population should be random to satisfy the property

  Example. the average of the sampling distribution of $\overline{y}$, $E(\overline{y})$, should be equal to the true value of the population mean, μ_y; if not, **bias** $= E(\overline{y}) - \mu_y \neq 0$

- **Consistency**: an estimator is **consistent** if it gets closer, as the sample size grows, to the true value of the parameter that is being estimated; uncertainty about μ_y decreases as n increases

Statistical inference

Problem: we have only one sample! We have to *infer* as much information as possible from the one sample we have

Solution: we use the variation in the one sample available to approximate the sampling distribution of our estimator $\overline{y}$

Intuition: the larger the sample we draw, the better we can approximate mean and variance of the sampling distribution

We have two tools:

- **Law of Large Numbers**: when sample size $n \to \infty$ then $\overline{y} \to \mu_y$ and $s_y^2 \to \sigma_y^2$ ($\overline{y}$: sample mean; s_y^2: sample variance)
- **Central Limit Theorem (CLT)** + **Law of Large Numbers**: when sample size $n \to \infty$, then the sampling distribution of $\overline{y}$ can be approximated by a normal: $\overline{y} \sim N\left(\mu_y, \frac{\sigma_y^2}{n}\right)$

Hypothesis Testing

Statistics let us test hypotheses about the world around us formulated as yes/no questions.

Example. Is the mean level of trust in political parties, μ_y, in European countries equal to 2 (on a 0 - 10 scale)?

Questions like this create two mutually exclusive statements, only one of which can be true:

- **Null hypothesis**: baseline statement we believe to be true

  $H_0 : \mu_y = 2$
- **Alternative hypothesis**: statement that holds true if the null is not

  $H_1 : \mu_y \neq 2$

Problem for policy analysts: decide whether to accept H_0 or to reject H_0 (in favor of H_1) using our one (random) sample and computing $\bar{y}$

T-statistic

The t-statistic is a *standardized* form of the sample average

Example. Is the mean level of trust in political parties, μ_y, in European countries equal to 2?

$$t = \frac{\bar{y} - 2}{SE(\bar{y})}$$

$SE(\bar{y})$ is the standard error of $\bar{y}$, computed as:

$$SE(\bar{y}) = \frac{s_y}{\sqrt{n}}$$

```
. ttest trstprt = 2

One-sample t test
```

Variable	Obs	Mean	Std. err.	Std. dev.	[95% conf. interval]	
trstprt	36,988	3.653536	.0127535	2.452794	3.628539	3.678534

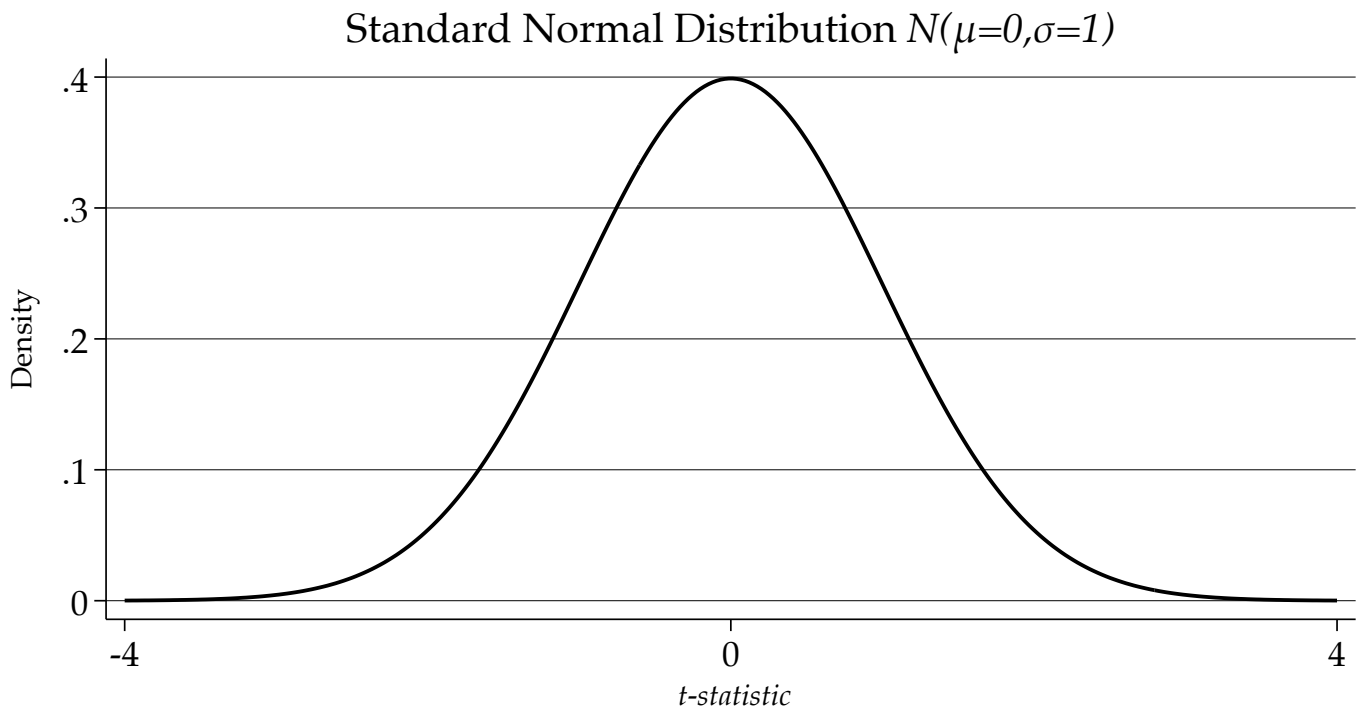
```
    mean = mean(trstprt)                                     t = 129.6532
H0: mean = 2                              Degrees of freedom =    36987

    Ha: mean < 2                 Ha: mean != 2                 Ha: mean > 2
 Pr(T < t) = 1.0000         Pr(|T| > |t|) = 0.0000          Pr(T > t) = 0.0000
```

T-statistic

How to interpret the t-statistic? Intuitively, the smaller (larger) $|t|$, the closer (farther) we are to (from) the value of H_0

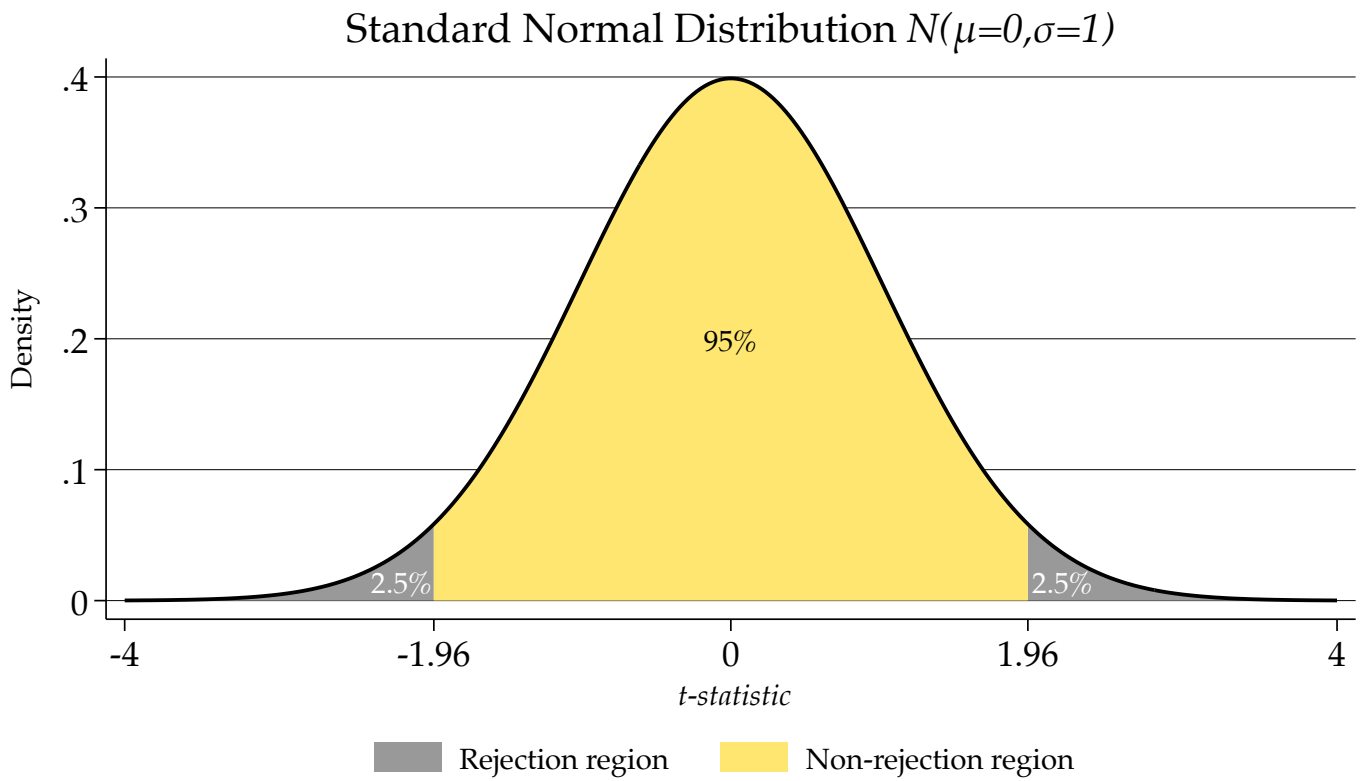
But, how large should be t to reject H_0? We don't want to make errors when taking this decision

By CLT, when $n \to \infty$, $t = \frac{\bar{y}-2}{SE(\bar{y})} \sim N(0,1)$



T-statistic

Decision rule to reject H_0: allow for a probability $\alpha = 5\%$, at most, to (incorrectly) reject H_0 when H_0 is true (α: *significance level* of a test)



5% corresponds to the area outside $[-1.96, +1.96]$ in a distribution $N(0,1)$ (1.96 being a *critical value*)

So, we reject H_0 if $|t| > 1.96$ with a $(1-\alpha)\% = 95\%$ confidence level

T-statistic

Example. Is the mean level of trust in political parties, μ_y, in European countries equal to **2**?

```
. ttest trstprt = 2

One-sample t test
```

Variable	Obs	Mean	Std. err.	Std. dev.	[95% conf. interval]	
trstprt	**36,988**	**3.653536**	**.0127535**	**2.452794**	**3.628539**	**3.678534**

```
    mean = mean(trstprt)                                     t = 129.6532
H0: mean = 2                              Degrees of freedom =    36987

    Ha: mean < 2                 Ha: mean != 2                 Ha: mean > 2
 Pr(T < t) = 1.0000         Pr(|T| > |t|) = 0.0000          Pr(T > t) = 0.0000
```

$$|t| = 129.65 > 1.96 \Rightarrow \textbf{Reject } H_0$$

Conclusion: political trust in European countries is different from 2 with a 95% level of confidence

T-statistic

Example. Is the mean level of trust in political parties, μ_y, in European countries equal to **3.65**?

```
. ttest trstprt = 3.65

One-sample t test
```

Variable	Obs	Mean	Std. err.	Std. dev.	[95% conf.	interval]
trstprt	36,988	3.653536	.0127535	2.452794	3.628539	3.678534

```
    mean = mean(trstprt)                                     t =   0.2773
H0: mean = 3.65                               Degrees of freedom =    36987

   Ha: mean < 3.65               Ha: mean != 3.65                 Ha: mean > 3.65
 Pr(T < t) = 0.6092         Pr(|T| > |t|) = 0.7816           Pr(T > t) = 0.3908
```

$$|t| = 0.28 < 1.96 \Rightarrow \textbf{Not reject } H_0$$

Conclusion: political trust in European countries is equal to 3.65 with a 95% level of confidence

Confidence intervals

Because we have a random sample, it is impossible to know the true population mean, μ_y, which is estimated by the sample mean, $\overline{y}$. But how accurate is our estimate?

Confidence Interval: range of values that contains the true population mean with a certain level of confidence (e.g. 95%)

$$CI_{0.95} = \overline{y} \pm 1.96 \cdot SE\left(\overline{y}\right) = [3.628; 3.678]$$

```
. ttest trstprt = 2

One-sample t test
```

Variable	Obs	Mean	Std. err.	Std. dev.	[95% conf. interval]	
trstprt	36,988	3.653536	.0127535	2.452794	3.628539	3.678534

```
    mean = mean(trstprt)                                      t = 129.6532
H0: mean = 2                               Degrees of freedom =    36987

    Ha: mean < 2                 Ha: mean != 2                 Ha: mean > 2
 Pr(T < t) = 1.0000         Pr(|T| > |t|) = 0.0000          Pr(T > t) = 0.0000
```

Confidence intervals

What if we require a more demanding test for the same hypotheses?

- probability of making a wrong decision, as well as significance level of the test, decreases (e.g. $\alpha = 1\%$)
- larger critical value in $N(0,1)$ (e.g. 2.58 for $\alpha = 1\%$)
- more difficult to reject (same) H_0
- confidence level of the decision increases (e.g. $(1-\alpha) = 99\%$)
- wider confidence interval containing the true mean (e.g. $CI_{0.99} = \overline{y} \pm 2.58 \cdot SE(\overline{y})$)

The opposite is true if we require a less demanding test (e.g. $\alpha = 10\%$)

Test for equality of means in two samples

Example. Is the mean level of trust in political parties in Italy, μ_{IT}, equal to that in the Netherlands, μ_{NL}?

Hypotheses: $H_0 : \mu_{IT} = \mu_{NL}$ and $H_0 : \mu_{IT} \neq \mu_{NL}$

Test statistic: $t = \frac{\bar{y}_{IT} - \bar{y}_{NL}}{SE(\bar{y}_{IT} - \bar{y}_{NL})} = \frac{3.122 - 5.348}{0.069} = -32.136$

```
. ttest trstprt if cntry=="IT"|cntry=="NL", by(cntry)

Two-sample t test with equal variances
```

Group	Obs	Mean	Std. err.	Std. dev.	[95% conf.	interval]
IT	2,606	3.122026	.0444503	2.269145	3.034865	3.209188
NL	1,454	5.348006	.0474562	1.80957	5.254915	5.441096
Combined	4,060	3.919212	.0371922	2.369818	3.846295	3.992129
diff		-2.225979	.0692673		-2.361781	-2.090177

```
    diff = mean(IT) - mean(NL)                                t = -32.1361
H0: diff = 0                                     Degrees of freedom =     4058

    Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
 Pr(T < t) = 0.0000         Pr(|T| > |t|) = 0.0000          Pr(T > t) = 1.0000
```

$$|t| = 32.13 > 1.96 \Rightarrow \textbf{Reject } H_0$$