Statistics Review

Sample Space and Events

The outcome of a random experiment cannot
be predicted with 100% accuracy. We need to assign probabilities to the possible outcomes of the experiment

Sample Space

The sample space of a random experiment is a set that contains every possible outcome of the experiment.

Event

An event is a subset of the sample space

Example

Suppose that we flip an unbiased coin twice. The sample space will be {HH, TT, HT, TH}. An event can be “obtaining heads on both flips” , which is {HH} .

Probability

For a sample space S, the probability measure **P** is a function that assigns each event E ⊂ S a number **P(**E) satisfying the following axioms:

1. 0$\leq$ P(E) ≤ 1

2. P(S) = 1

3. For any sequence of mutually exclusive events E1, E2, ... (that is, events for which $E_i \cap E_j$ = ∅ when i ≠ j), we have

$$\boldsymbol{P}\left[\underset{i \geq 1}{\cup}\left(E_{i}\right)\right] = \sum_{\mathrm{i} \geq 1} P\left[E_{i}\right]$$

Axioms of Probability

Let **P** be a probability measure.

(1) (Monotonicity.) If A ⊂ B, then **P**[A] ≤ **P**[B].

(2) (Complement rule.) For every event E, one has **P**[E^c] = 1 − **P**[E].

(3) (Empty event.) **P**[∅] = 0

(4) (Inclusion-exclusion formula.) For every events A, B ⊂ S, one has
P[A ∪ B] = **P**[A] + **P**[B] − **P**[A ∩ B]

Conditional Probability

Let S be a sample space.
Suppose a random event B is drawn

The conditional probability of another event A given B is the probability that A happened given that B happened.

Quantitatively, it is $\mathbf{P}(A \mid B) = \frac{\mathbf{P}(A \cap B)}{\boldsymbol{P}(B)}$

This is known as “the probability of A conditioned on B ”

Independent events

We say that two events A and B such that **P**(A), **P**(B) > 0 are independent if
$\mathbf{P}(A|B) = \mathbf{P}(A)$ and $\mathbf{P}(B|A) = \mathbf{P}(B)$. or, equivalently, if

$\mathbf{P}(A \cap B) = \mathbf{P}(A)\mathbf{P}(B)$

Intuitively: knowing A occurred does not change the likelihood that B occurred, and vice versa.

Example: Flipping a coin

Random Variable

- A **random variable**, usually denoted by X, is a rule that assigns a numerical value to each outcome in a sample space. Random variables may be either discrete or continuous.
- We use random variables (r.v.) to model data that are uncertain, e.g.

Number of heads in ten coin tosses

Share of votes for a candidate in an election

Average # of hours spent on homework each week

Household income in the U.S.

Indicator Functions

An Important application of random variable is an indicator function:

I{E} = $\begin{cases} 1 \; if \; E \; occurs \\ 0 \; otherwise \end{cases}$

Expected Value

Expected value, or **expectation**/population mean, is the weighted average of the possible values that the variable can take, weighed on the probability of each value occurring.

$$\boldsymbol{E}[X] = \sum_{x \in R_x} x \, P\left[X = x\right]$$

Example:

If a random variable X takes on values of -1, 0, and 2, with probabilities 0.3, 0.3, and 0.4 respectively, then the expectation of X equals

$$\boldsymbol{E}[X] = (-1)(0.3) + (0)(0.3) + (2)(0.4) = 0.5$$

Linearity of Expectation

Let X and Y be two discrete random variables and let $a \in \mathbb{R}$ be a nonrandom constant. Then,

$$\mathbf{E}[X + Y] = \mathbf{E}[X] + \mathbf{E}[Y]$$

and

$$\mathbf{E}[aX] = a\mathbf{E}[X]$$

Variance

Suppose that X has expectation E[X]. Its variance is
$\mathrm{Var}(X) = E[(X - E[X]^2)] = E[X^2] - E[X]^2$

The variance is a measure of how far X will be, on average, from its expected value E[X]. Stated another way, the variance measures how random a random variable is.

Note: Standard Deviation SD[X] = $\sqrt{Var[X]}$

Conditional Expectation

How the expected value of a random variable gets updated once we observe that some event occurs.

The expectation of a random variable X conditional on some event A is denoted by E[X|A] .

$$\boldsymbol{E}\left[X|A\right] = \sum_{x \,\in\, Outcomes\ of\ \ X\ given\ A} x\ \boldsymbol{P}\left[X = x|A\right]$$

Linearity with Conditioning

Let X and Y be discrete random variables and *a* be a constant. A is an event on the sample space. Then,

$\mathbf{E}[X + Y | A] = \mathbf{E}[X|A] + \mathbf{E}[Y|A]$

and

$\mathbf{E}[aX|A] = a\mathbf{E}[X|A]$

Covariance

- The covariance measures the linear dependence between two random variables
- Covariance between X and Y is **Cov**(X,Y) = **E**[(X– **E**[X])(Y-**E**[Y])]
- Positive: Y increases as X increases
- Negative: Y decreases as X increases

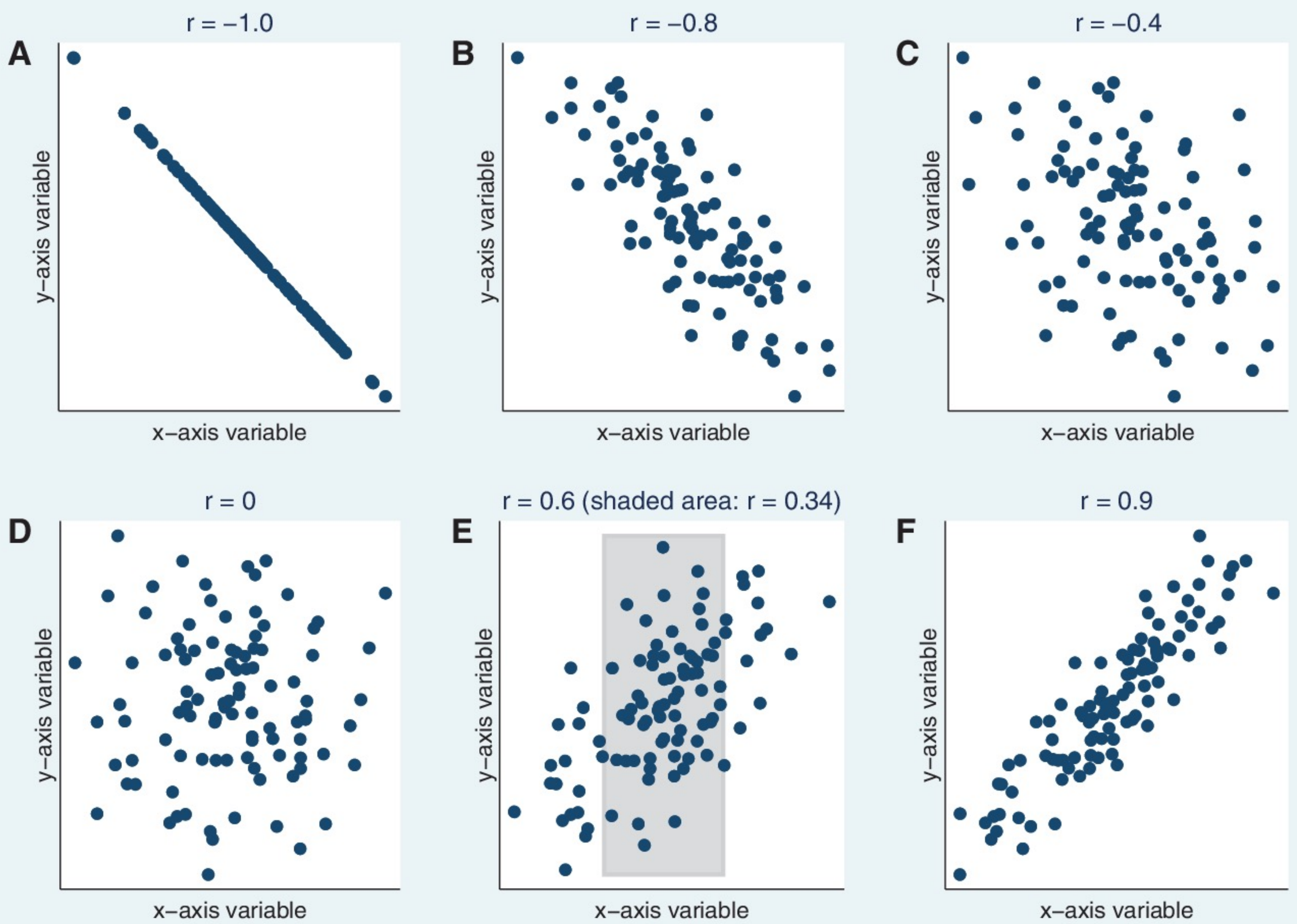
Correlation

Given a pair of random variables (X,Y) , the correlation coefficient ρ is defined as

$$\rho_{X,Y} = \frac{cov(X,Y)}{SD[X]SD[Y]}$$

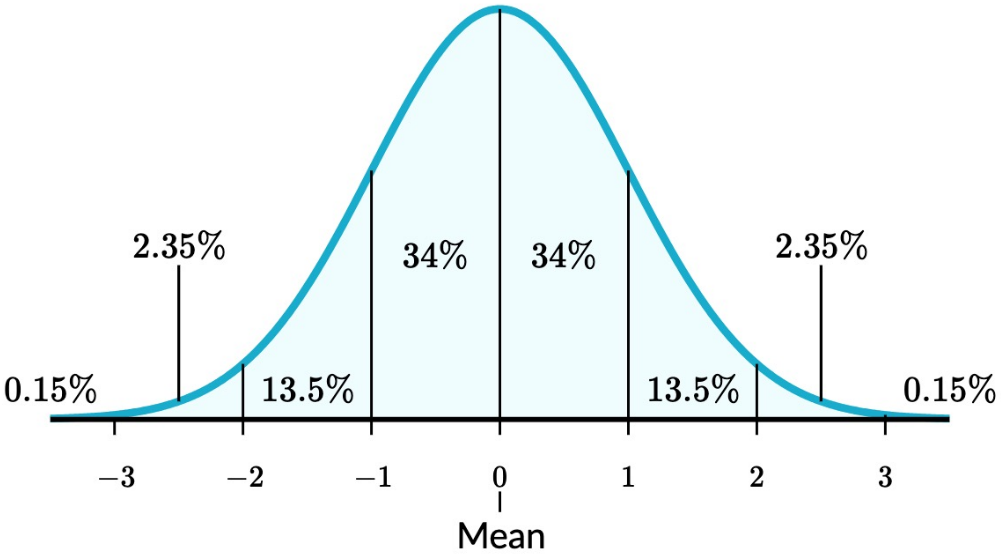
$$-1 \leq \rho_{X,Y} \leq 1$$

Correlation measures the strength of the relationship between X and Y


r = −1.0
A
y−axis variable
x−axis variable
r = −0.8
B
y−axis variable
x−axis variable
r = −0.4
C
y−axis variable
x−axis variable
r = 0
D
y−axis variable
x−axis variable
r = 0.6 (shaded area: r = 0.34)
E
y−axis variable
x−axis variable
r = 0.9
F
y−axis variable
x−axis variable

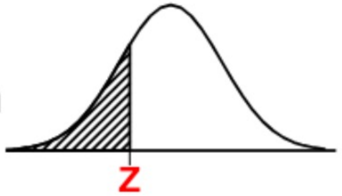
Standard Normal Distribution

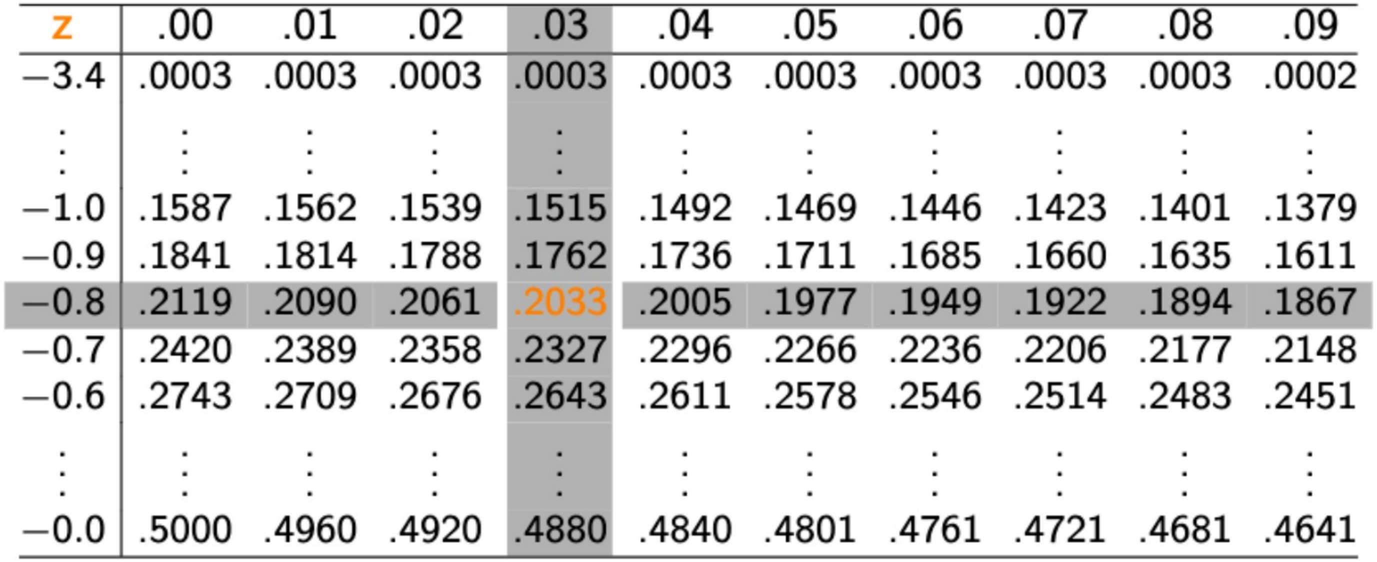
A random variable X is said to have a standard normal distribution if its normal distribution has μ (mean) = 0 and $\sigma(standard\ deviation) = 1$



Standard Normal Distribution Table

$\Phi(z) = P(Z < z) =$ area of shaded region in





z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
−3.4	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0002
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
−1.0	.1587	.1562	.1539	.1515	.1492	.1469	.1446	.1423	.1401	.1379
−0.9	.1841	.1814	.1788	.1762	.1736	.1711	.1685	.1660	.1635	.1611
−0.8	.2119	.2090	.2061	.2033	.2005	.1977	.1949	.1922	.1894	.1867
−0.7	.2420	.2389	.2358	.2327	.2296	.2266	.2236	.2206	.2177	.2148
−0.6	.2743	.2709	.2676	.2643	.2611	.2578	.2546	.2514	.2483	.2451
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
−0.0	.5000	.4960	.4920	.4880	.4840	.4801	.4761	.4721	.4681	.4641

E.g., for z = −0.83, find Φ(z).

Z-score

Z-score measures the number of standard deviations by which a raw data point is above or below the observed mean.

$$z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

High level overview of Statistics

- We use statistics to say something about population-level patterns using a sample of data
  - For example, we may care about average height in a country
  - In practice, we cannot survey every household so we survey a sample of them

There are three major things we can do:
1. Estimate a population object (e.g. the average height of a population)
2. Test a hypothesis (e.g. we believe average height is 5’10)
3. Create a confidence interval (e.g. we have an interval that will contain the true average height 95% of the time)

Hypothesis Tests

If we want to test a hypothesis, we propose a “null” and an “alternative” hypothesis.

The average height of all residents in town X is recorded to be 168 cm. A scientist believes that the true mean is smaller. She measured the height of 36 individuals and found the mean to be 165 cm with a standard deviation of 5.

Hypothesis Tests

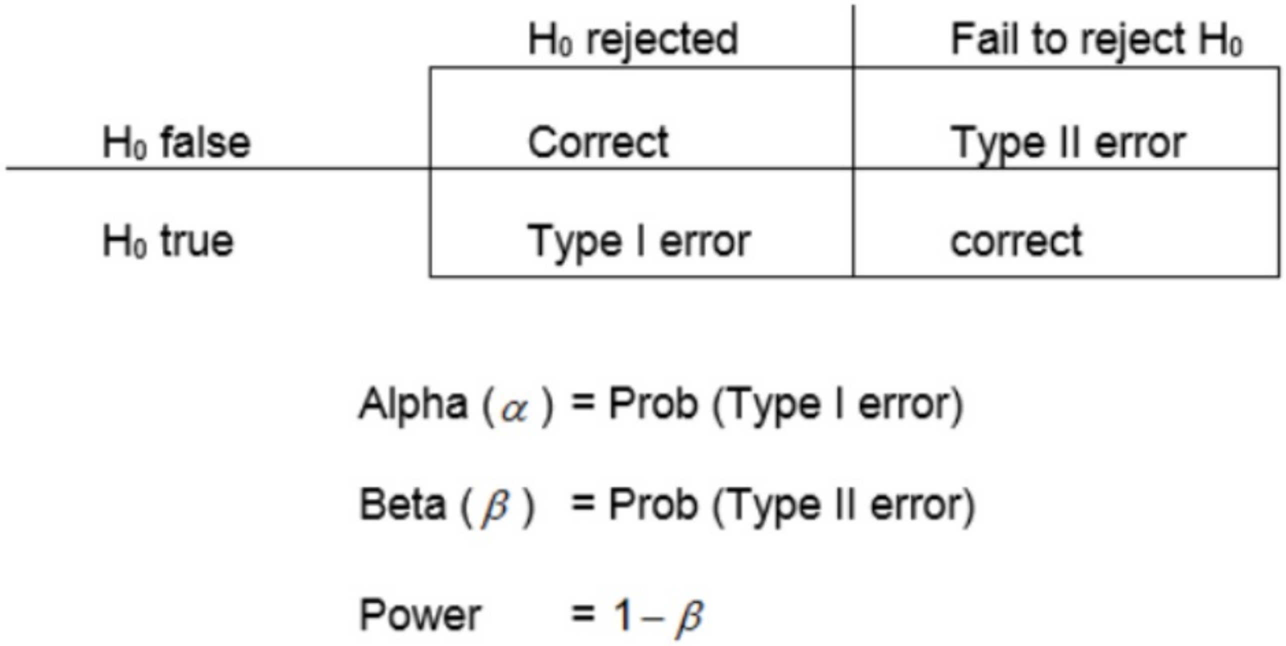
The average height of all residents in town X is recorded to be 168 cm. A scientist believes that the true mean is smaller. She measured the height of 36 individuals and found the mean to be 165 cm with a standard deviation of 5. At 5% significance level, is there enough evidence to support her beliefs?

Null hypothesis : $H_0 : \mu = 168$

Alternative Hypothesis: $H_1 : \mu < 168$

Then we want to minimize Type I and Type II error.

Type I and Type II Error

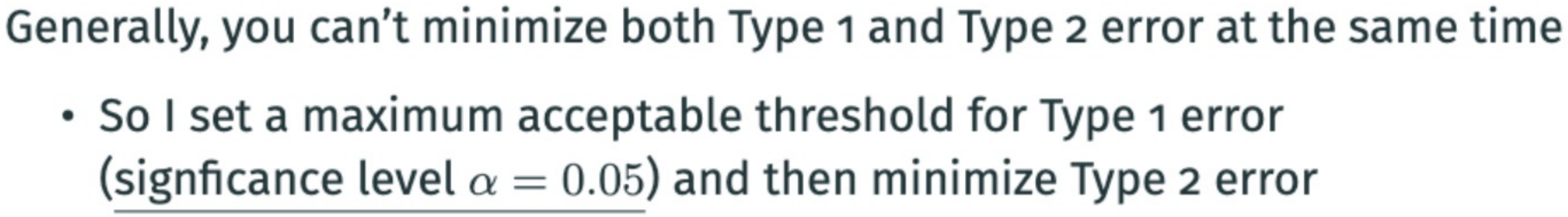


	H_0 rejected	Fail to reject H_0
H_0 false	Correct	Type II error
H_0 true	Type I error	correct

Alpha (α) = Prob (Type I error)

Beta (β) = Prob (Type II error)

Power = $1 - \beta$



Generally, you can’t minimize both Type 1 and Type 2 error at the same time

- So I set a maximum acceptable threshold for Type 1 error (signficance level $\alpha = 0.05$) and then minimize Type 2 error

P-value

If the null hypothesis was assumed to be true, the **p-value** is the probability of obtaining test results at least as extreme as the value observed.

In other words, if we assume that the null hypothesis was true, what is the probability that we got this result from the sample ?

Rule: reject null hypothesis at significance level α if p-value is less than α.

Hypothesis Tests

The average height of all residents in town X is 168 cm. A scientist believes that the true mean is smaller. She measured the height of 36 individuals and found the mean to be 165 cm with a standard deviation of 5. At 5% significance level, is there enough evidence to support her beliefs?

Null hypothesis : $H_0 : \mu = 168$

Alternative Hypothesis: $H_1 : \mu < 168$

Hypothesis Tests

The average height of all residents in town X is 168 cm. A scientist believes that the true mean is smaller. She measured the height of 36 individuals and found the mean to be 165 cm with a standard deviation of 5. At 5% significance level, is there enough evidence to support her beliefs?

Null hypothesis : $H_0 : \mu = 168$

Alternative Hypothesis: $H_1 : \mu < 168$

Hypothesis Tests

The average height of all residents in town X is 168 cm. A scientist believes that the true mean is smaller. She measured the height of 36 individuals and found the mean to be 165 cm with a standard deviation of 5. At 5% significance level, is there enough evidence to support her beliefs?

Null hypothesis : $H_0 : \mu = 168$

Alternative Hypothesis: $H_1 : \mu < 168$

STANDARD NORMAL DISTRIBUTION: Table Values Represent AREA to the LEFT of the Z score.

Z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
-3.9	.00005	.00005	.00004	.00004	.00004	.00004	.00004	.00004	.00003	.00003
-3.8	.00007	.00007	.00007	.00006	.00006	.00006	.00006	.00005	.00005	.00005
-3.7	.00011	.00010	.00010	.00010	.00009	.00009	.00008	.00008	.00008	.00008
-3.6	.00016	.00015	.00015	.00014	.00014	.00013	.00013	.00012	.00012	.00011
-3.5	.00023	.00022	.00022	.00021	.00020	.00019	.00019	.00018	.00017	.00017
-3.4	.00034	.00032	.00031	.00030	.00029	.00028	.00027	.00026	.00025	.00024
-3.3	.00048	.00047	.00045	.00043	.00042	.00040	.00039	.00038	.00036	.00035
-3.2	.00069	.00066	.00064	.00062	.00060	.00058	.00056	.00054	.00052	.00050
-3.1	.00097	.00094	.00090	.00087	.00084	.00082	.00079	.00076	.00074	.00071
-3.0	.00135	.00131	.00126	.00122	.00118	.00114	.00111	.00107	.00104	.00100
-2.9	.00187	.00181	.00175	.00169	.00164	.00159	.00154	.00149	.00144	.00139
-2.8	.00256	.00248	.00240	.00233	.00226	.00219	.00212	.00205	.00199	.00193
-2.7	.00347	.00336	.00326	.00317	.00307	.00298	.00289	.00280	.00272	.00264
-2.6	.00466	.00453	.00440	.00427	.00415	.00402	.00391	.00379	.00368	.00357
-2.5	.00621	.00604	.00587	.00570	.00554	.00539	.00523	.00508	.00494	.00480
-2.4	.00820	.00798	.00776	.00755	.00734	.00714	.00695	.00676	.00657	.00639