Chapter 8
Hypothesis Testing
And
Linear Regression

Hypothesis Testing

A statistical hypothesis is an assertion or conjecture concerning one or more populations.

To prove that a hypothesis is true with absolute certainty (or to prove that it is false with absolute certainty), we would have to examine the entire population.

Instead, take a random sample from the population and use it as evidence that either supports or does not support the hypothesis.

H_0 = Null hypothesis. The hypothesis we wish to test

The rejection of H_0 leads to the acceptance of an alternate hypothesis H_1. 2 possible outcomes:

1. reject H_0: in favor of H_1 because of sufficient evidence in the sample.
2. fail to reject H_0: because of insufficient evidence

Note: There is no formal outcome that says accept H_0. H_0 often represents "status quo" in opposition to a new idea H_1.

Example: Jury Trial H_0: defendant is innocent
H_1: " is guilty

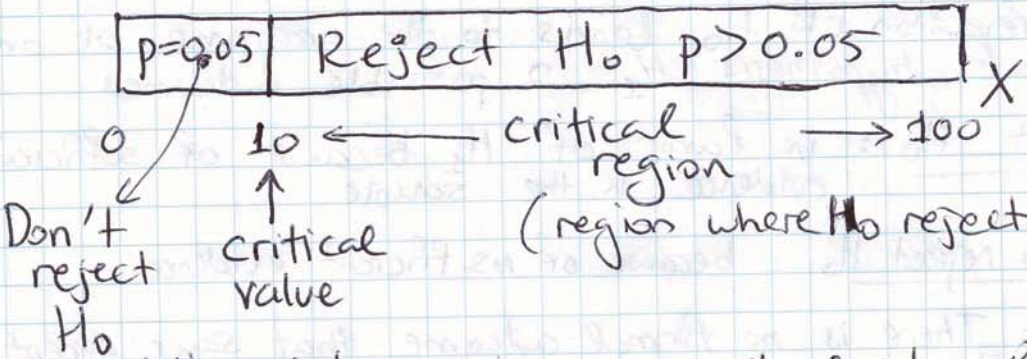
H_0 is rejected if H_1 is supported by evidence beyond a reasonable doubt. Failure to reject H_0 doesn't imply innocence, only that the evidence was not sufficient to reject it. Therefore, failure to reject H_0 does not necessarily mean accept H_0.

Example: A company manufacturing RAM chips claims the defective rate of the population is 5%. Let p denote the true defective probability

$H_0 : p = 0.05$
$H_1 : p > 0.05$

} Take a sample of 100 RAM chips off the production line and test.

Let X denote the number of defectives in the sample of 100. Reject H_0 if $X \geq 10$
(Note 10 is chosen arbitrarily in this case)
X is called the test statistic



Q: Why did we choose a critical value of 10 in this example?

A: Since we know this is a Bernoulli Process, the expected value of defectives is np. So if we believe that $p = 0.05$, we should expect $100 \times 0.05 = 5$ defectives in a sample of 100. Therefore, 10 defectives would be strong evidence that $p > 0.05$. Later we will learn how to choose the critical value based on the desired level of significance for the hypothesis test.

Possible situations

	H_0 is true	H_0 is false
Do not reject H_0	Correct decision	Type II error
Reject H_0	Type I error	Correct decision

Defn: Rejection of H_0 when it is true is called a Type I error.

Example: Convicting the defendant when he is innocent.

Defn: The probability of committing a type I error, denoted by α, is called the level of significance.

Example (RAM chips continued)

$$\alpha = P(\text{Type 1 error}) = P(\underbrace{X \geq 10}_{\text{critical region}} \text{ when } \underbrace{p = 0.05}_{H_0 \text{ is true.}})$$

$$= \sum_{x=10}^{100} b(x; n=100, p=0.05) \quad \text{Binomial Distribution}$$

$$= \sum_{x=10}^{100} \binom{100}{x} 0.05^x \, 0.95^{100-x} = 0.0282$$

Level of significance is $\alpha = 0.0282$.

The lower the α, the less likely we are to commit a type I error. Would like small α values. (0.05 or smaller generally used)

Defn: Nonrejection of H_0 when it is false is called a type II error. The probability of committing a type II error is denoted β.

Note: β is impossible to compute unless we have a specific alternate hypothesis.

Example continued: Can't compute β for $H_1: p > 0.05$ but can compute it for testing $H_0: p = 0.05$ against the alternate hypothesis that $p = 0.1$ for instance.

$$\beta = P(\text{Type II error}) = P(X < 10 \text{ when } p = 0.1)$$

$$= \sum_{x=0}^{9} b(x; n = 100, p = 0.1) = 0.4513$$

$\uparrow$ H_1 (H_0 false)

This high probability suggests we are likely to fail to reject H_0 if the true p is 0.1. This might be OK if we only want to make sure we don't fail to reject if the true p is 0.15

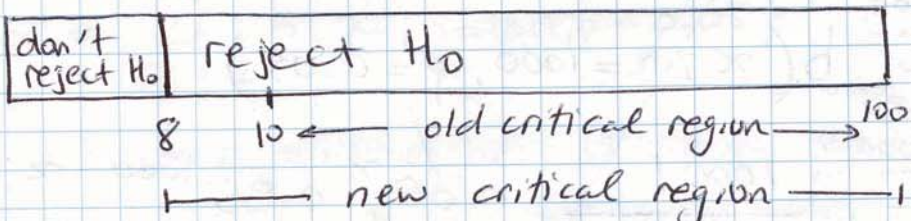
$$\beta = P(\text{Type II error}) = P(X < 10 \text{ when } p = 0.15)$$

$$= \sum_{x=0}^{9} b(x; n = 100, p = 0.15) = 0.0551$$

Effect of critical value

Moving the critical value provides a trade-off between α and β. A reduction in β is always possible by increasing the size of the critical region, but this increases α.

Example continued: Lets change the critical value from 10 to 8. So reject H_0 if $X \geq 8$



$$\alpha = \sum_{x=8}^{100} b(x; n=100, p=0.05)$$ Probability of rejecting H_0 when it is true

$$= 0.128$$ (note larger than before when the critical value was set at 10)

testing against alternate hypothesis $p = 0.1$

$$\beta = \sum_{x=0}^{7} b(x; n=100, p=0.1)$$ Probability of not rejecting H_0 when $p = 0.1$

$$= 0.206$$ (lower than before)

testing against alternate hypothesis $p = 0.15$

$$\beta = \sum_{x=0}^{7} b(x; n=100, p=0.15)$$ Probability of not rejecting H_0 when $p = 0.15$

$$= 0.012$$ (again lower than before)

Effect of sample size

Both α and β can be reduced simultaneasly by increasing the sample size.

Example continued : Sample size $n = 150$, critical value 12. Reject H_0 if $X \geq 12$ (X is the # of defectives in sample of 150)

$$\alpha = \sum_{x=12}^{150} b(x; n=150, p=0.05) = 0.074$$

(was 0.128 for $n=100$ and critical value 8)

testing against alternate hypothesis $p = 0.1$

$$\beta = \sum_{x=0}^{11} b(x; n=150, p=0.1) = 0.171$$

(was 0.206 for $n=100$ and critical value 8)

— o —

Factorials of very large numbers problematic to compute accurately even with MATLAB. Thankfully, the Binomial Distribution can be approximated by the normal distribution - Details Section 6.5

Theorem : If X is a binomial random variable with n trials and probability of success of each trial p, then the limiting form of the distribution of $Z = \dfrac{X - np}{\sqrt{np(1-p)}}$ as $n \to \infty$ is the standard normal distribution.

— o —

Approximation is good when n is large and p is not extremely close to 0 or 1.

Lets recompute our α with the normal approximation.

$$\alpha = P(\text{Type I error}) = \sum_{x=12}^{150} b(x; n=150, p=0.05)$$

$$\approx P\left(Z \geq \frac{12 - 150 \times 0.05}{\sqrt{150 \times 0.05 \times 0.95}}\right) = P(Z \geq 1.69)$$

$$= 1 - P(Z \leq 1.69) = 1 - 0.9545 = 0.0455$$

not too bad.

What if we increase the sample size to $n = 500$ and the critical value to 40?

The normal approximation should be even better since n larger

$$\alpha \approx P\left(Z \geq \frac{40 - 500 \times 0.05}{\sqrt{500 \times 0.05 \times 0.95}}\right) = P(Z \geq 3.08)$$

$$= 1 - P(Z \leq 3.08) = 1 - 0.999 = 0.001$$

Very unlikely to commit type I error.

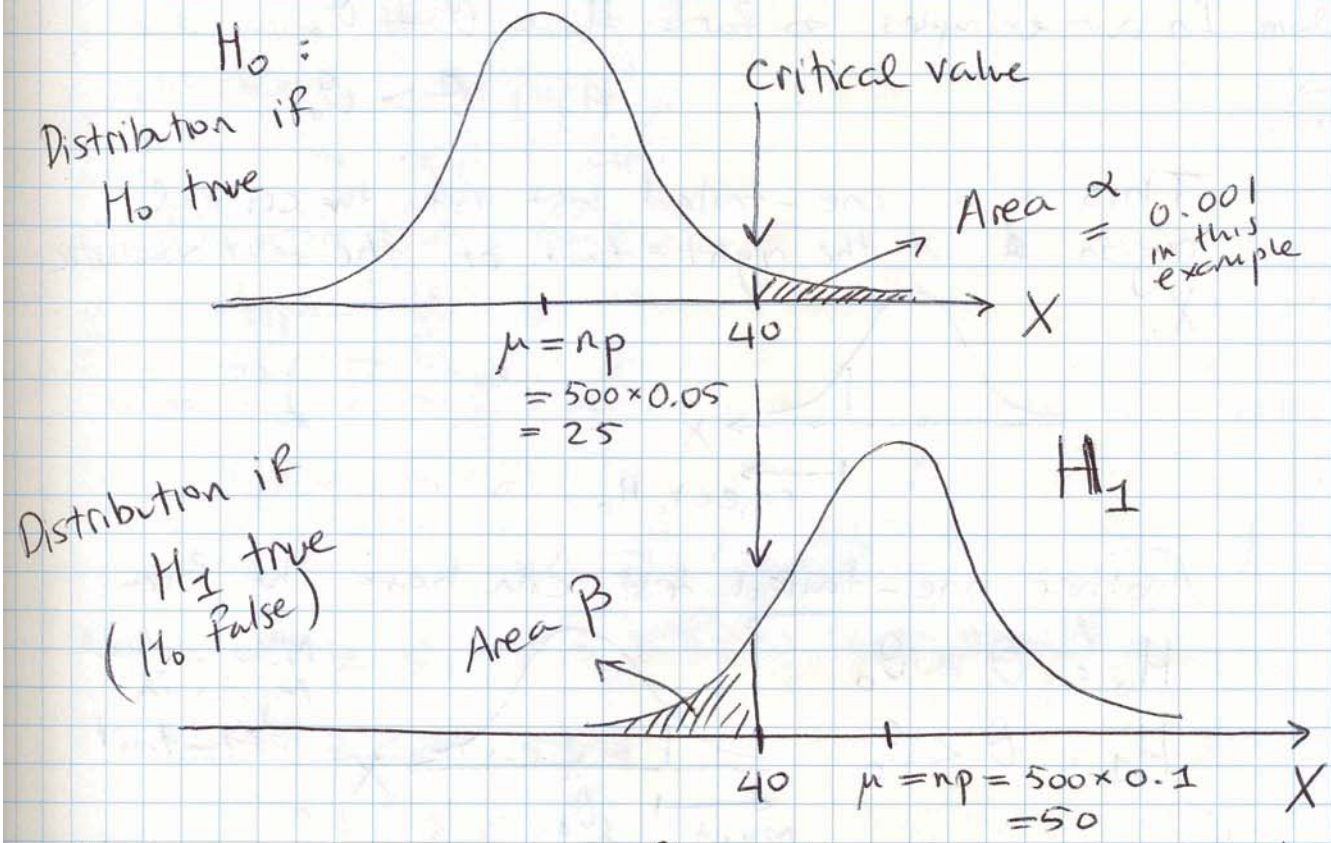
testing against alternate hypothesis $p = 0.1$

$$\beta = \sum_{x=0}^{39} b(x; n=500, p=0.1)$$

$$\approx P\left(Z \leq \frac{39 - 500 \times 0.1}{\sqrt{500 \times 0.1 \times 0.9}}\right) = P(Z \leq -1.49)$$

$$= 0.0681$$

Visual interpretation with normal approximation



Defn: The power of a test is the probability of rejecting H_0 given that a specific alternate is true. Power $= 1 - \beta$.

Properties of hypothesis testing

(1) α and β are related. Decreasing one generally increases the other

(2) α can be set to a desired value by adjusting the critical value. α typically set at 0.05 or 0.01

(3) Increasing n decreases α and β both.

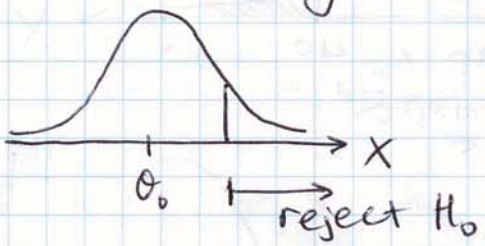
(4) β decreases as the distance between the true value and hypothesized (H_1) value increases.

One-tailed vs. two-tailed tests

In our examples so far $H_0: \theta = \theta_0$

$H_1: \theta > \theta_0$

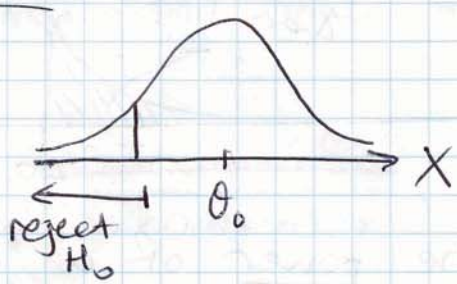
This is a one-tailed test with the critical region in the right-tail of the test statistic X.



Another one-tailed test can have the form

$H_0: \theta = \theta_0$

$H_1: \theta < \theta_0$

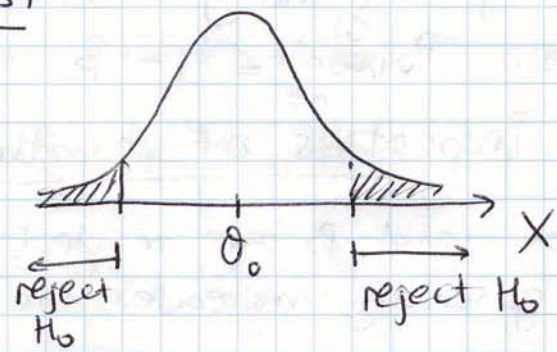


Now critical region in left-tail.

Two-sided test

$H_0: \theta = \theta_0$

$H_1: \theta \neq \theta_0$



Example: Production line of resistors that are supposed to be 100 ohms. Assume $\sigma = 8$

$H_0: \mu = 100$

$H_1: \mu \neq 100$

Let $\bar{X}$ be the sample mean for a sample of size $n = 100$

Reject H_0	Do not reject H_0	Reject H_0

98 102

The test statistic is the sample mean in this case. We know the sampling distribution of $\bar{X}$ is a normal distribution with mean μ and standard deviation $\sigma/\sqrt{n}$ due to the central limit theorem. $(n=100>30)$

Therefore we can compute the probability of a type I error as

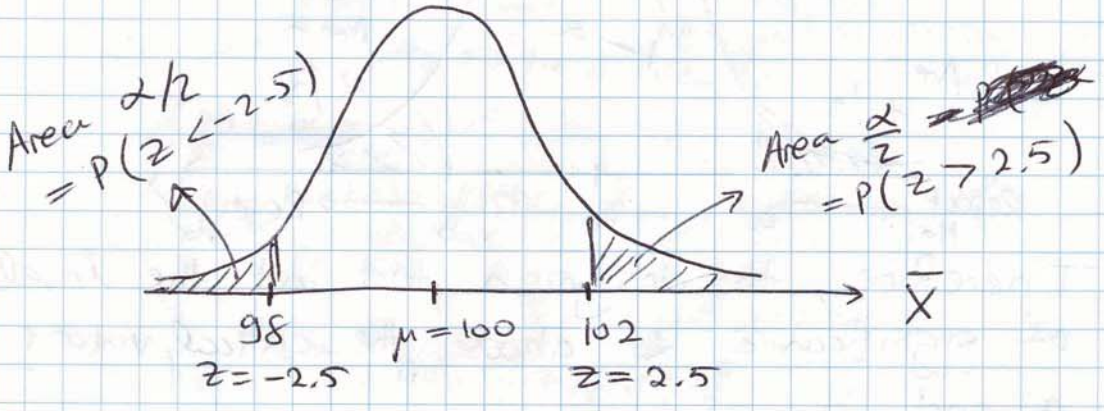
$$\alpha = P(\bar{X} < 98 \text{ when } \mu=100) + P(\bar{X} > 102 \text{ when } \mu=100)$$

$$= P\left(z < \frac{98-100}{8/\sqrt{100}}\right) + P\left(z > \frac{102-100}{8/\sqrt{100}}\right)$$

$$= P(z < -2.5) + P(z > 2.5)$$

$$= P(z < -2.5) + \left(1 - P(z < 2.5)\right)$$

$$= 2P(z < -2.5) = 2 \times 0.0062 = 0.0124$$



* Testing $H_0: \mu = \mu_0$ against $H_1: \mu \neq \mu_0$ at a significance level α is equivalent to computing a $100(1-\alpha)\%$ confidence interval for μ and rejecting H_0 if μ_0 is outside the confidence interval.

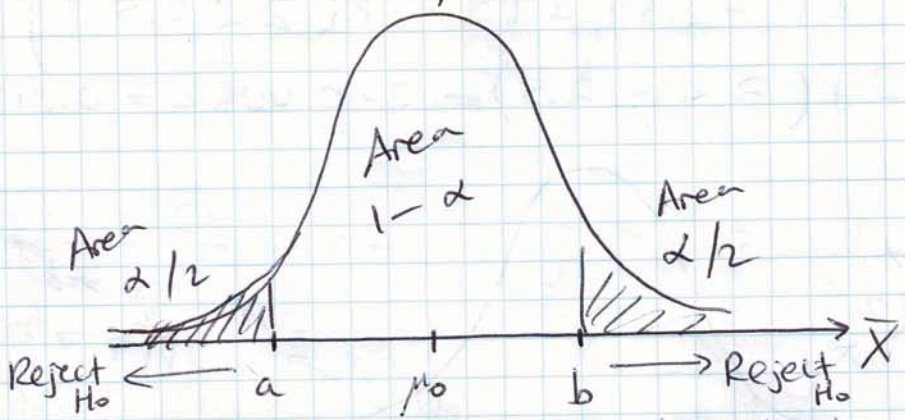
Tests concerning sample mean (Variance known)

$H_0 : \mu = \mu_0$
$H_1 : \mu \neq \mu_0$

Sample $X_1, \dots, X_n$
Sample mean $\bar{X}$
Known population variance σ^2

Under H_0 $\mu = \mu_0$ so P(type 1 error) is computed using the sampling distribution of $\bar{X}$ which is normal due to the central limit theorem with mean μ and standard deviation $\sigma/\sqrt{n}$. From confidence intervals we know that

$$P\left(-z_{\alpha/2} < \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} < z_{\alpha/2}\right) = 1 - \alpha$$



Therefore, to design a test at the level of significance α choose the critical values a and b as

$$a = \mu_0 - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

$$b = \mu_0 + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

Then collect sample, compute the sample mean $\bar{X}$ reject H_0 if $\bar{X} < a$ OR $\bar{X} > b$

Steps in hypothesis testing

1. State the null and alternative hypothesis
2. Choose a significance level α
3. Choose test statistic and establish critical region
4. Collect sample, compute test statistic on sample
   Reject H_0 if test statistic in critical region.
   Otherwise do not reject H_0

Example: A batch of 100 resistors have an average of 102 Ohms. Assuming a population standard deviation of 8 Ohms. Test whether the population mean is 100 Ohms at a significance level $\alpha = 0.05$.

Soln: (1) $H_0 : \mu = \underbrace{100}_{\text{This is } \mu_0}$ $H_1 : \mu \neq 100$
(Note unless otherwise stated we use a two-tailed test)

(2) $\alpha = 0.05$

(3) Test statistic $\bar{X}$, Reject H_0 if
$\bar{X} < a$ or $\bar{X} > b$

$z_{\alpha/2} = z_{0.025} = 1.96$

$$a = \mu_0 - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} = 100 - 1.96\frac{8}{10}$$
$$= 98.432$$

$$b = \mu_0 + z_{\alpha/2}\frac{\sigma}{\sqrt{n}} = 100 + 1.96\frac{8}{10}$$
$$= 101.568$$

(4) $\bar{X} = 102 > b$ therefore
Reject H_0.

One-sided tests of the sample mean

(A) $H_0: \mu = \mu_0$ Reject H_0 at significance level α
$H_1: \mu > \mu_0$ if $\bar{X} > \mu_0 + z_\alpha \frac{\sigma}{\sqrt{n}}$

↑ Note that z_α appears instead of $z_{\alpha/2}$ just like in one-tailed confidence intervals

(B) $H_0: \mu = \mu_0$ Reject H_0 at a significance level
$H_1: \mu < \mu_0$ α if $\bar{X} < \mu_0 - z_\alpha \frac{\sigma}{\sqrt{n}}$

Example: A quality control engineer finds that a sample of 100 light bulbs had an average life-time of 470 hours. Assuming a population standard deviation $\sigma = 25$ hours, test whether the population mean is 480 hours vs. the alternative hypothesis $\mu < 480$ at a significance level $\alpha = 0.05$

Soln: ① $H_0: \mu = 480$ $H_1: \mu < 480$ (one-tailed test)

② $\alpha = 0.05$

③ Test statistic $\bar{X}$. Reject H_0 if

$$\bar{X} < \mu_0 - z_\alpha \frac{\sigma}{\sqrt{n}} = 480 - 1.645 \frac{25}{10}$$

$$\bar{X} < 475.9$$

④ Since $\bar{X} = 470$ Reject H_0

Note: Table A.3 has no entry with $P(Z < z) = 0.05$ exactly. The closest ones are $P(Z < -1.64) = 0.505$ and $P(Z < -1.65) = 0.495$

The book (and in this example we) use the average of 1.64 and 1.65 as the z-value for which $P(Z > z) = 0.05$. Alternatively, we could just use whichever is closest to the α value we are looking for.

Tests concerning sample mean (Variance unknown)

$H_0 : \mu = \mu_0$
$H_1 : \mu \neq \mu_0$

} Sample $X_1, \ldots, X_n$ from a normal population. Unknown σ^2.
Sample mean $\bar{X}$, sample variance S^2

We know that in this case the sampling distribution for $\bar{X}$ is the t-distribution.

Critical region at significance level α is

$\bar{X} < a$ OR $\bar{X} > b$ reject H_0

where $a = \mu_0 - t_{\alpha/2} \frac{S}{\sqrt{n}}$

$b = \mu_0 + t_{\alpha/2} \frac{S}{\sqrt{n}}$

where $t_{\alpha/2}$ is from row with $\nu = n-1$

OR equivalently let $T = \frac{\bar{X} - \mu_0}{S/\sqrt{n}}$

reject H_0 if $T < -t_{\alpha/2}$ or $T > t_{\alpha/2}$ for $\nu = n-1$ degrees of freedom.

In one-sided test $t_{\alpha/2}$ is replaced by t_α as usual.

Example (Example 10.5 textbook)

Claim: A certain electrical appliance (vacuum cleaner) expends 46 kWh per year. A random sample of 12 homes indicate that vacuum cleaners expend an average of 42 kWh per year with standard deviation 11.9 kWh. Does this suggest at the 0.05 level of significance that, on the average, vacuum cleaners expend less than 46 kWh per year? Assume population of kWh to be normally distributed.

Soln: (1) $H_0 : \mu = 46$ kWh

$H_1 : \mu < 46$ kWh

(2) $\alpha = 0.05$

(3) Test statistic

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

reject H_0 if

$T < -t_{0.05}$ for $v = n - 1 = 11$ degrees of freedom.

$t_{0.05} = 1.796$

so reject if $T < -1.796$

(4) $\bar{X} = 42 \quad s = 11.9 \quad n = 12$

$$T = \frac{42 - 46}{11.9/\sqrt{12}} = -1.16$$

Do not reject H_0

P-values

Preselection of significance level α has its roots in the philosophy that making a type-I error should be controlled.

The P-value approach aims to give more information to the user, especially when the test-statistic is close to the critical region.

Defn: A P-value is the lowest level of significance at which the observed value of a test statistic is significant (rejects H_0).

Example: A batch of 100 resistors have an average of 101.5 Ohms. Assuming a population standard deviation of 5 Ohms

a) Test whether the population mean is 100 Ohms at level of significance $\alpha = 0.05$

b) Compute p-value.

Soln a) $H_0 : \mu = 100 \quad H_1 : \mu \neq 100$

Test statistic $\bar{X}$. Reject H_0 if

$$\bar{X} < 100 - z_{0.025}\frac{\sigma}{\sqrt{n}} = 100 - 1.96 \times \frac{5}{10} = 99.02$$

OR

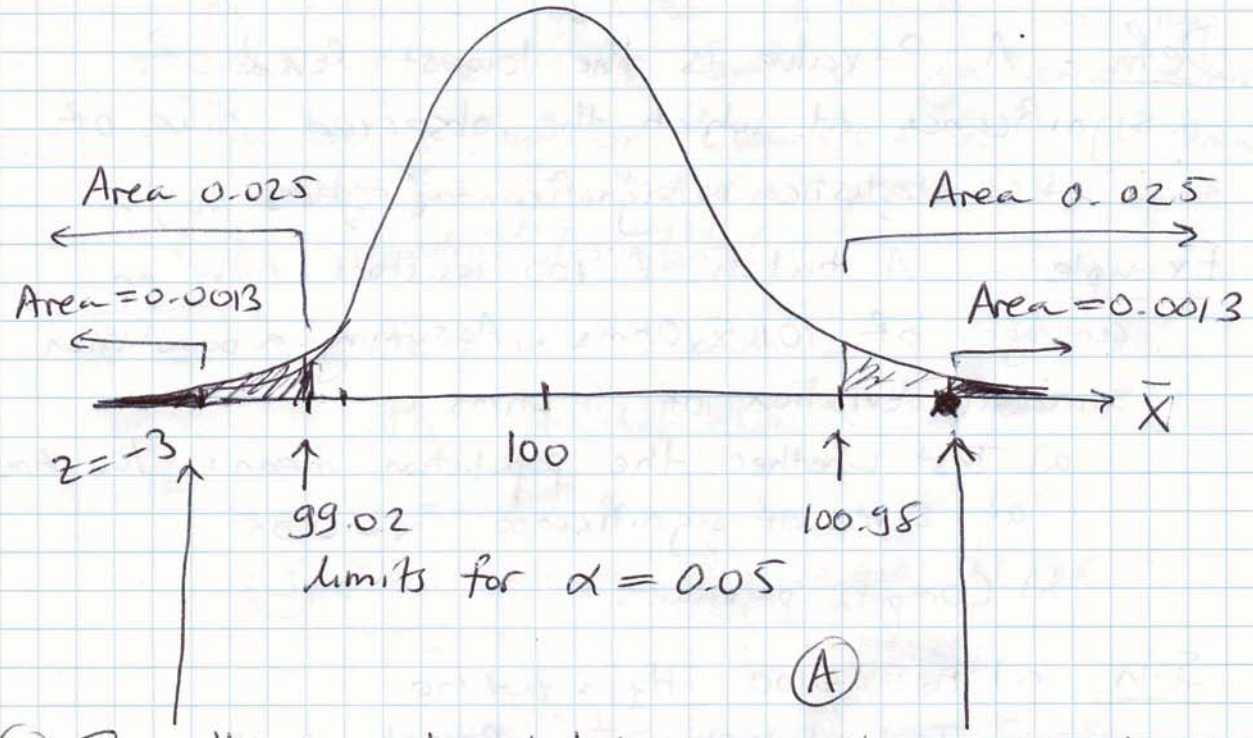
$$\bar{X} > 100 + z_{0.025}\frac{\sigma}{\sqrt{n}} = 100 + 1.96 \times \frac{5}{10} = 100.98$$

$\bar{X} = 101.5$ therefore reject H_0

b) $$z = \frac{\bar{X} - 100}{\sigma/\sqrt{n}} = \frac{101.5 - 100}{5/10} = 3$$ observed z-value

$$P = 2P(Z > 3 \text{ when } \mu = 100) = 2 \times 0.0013 = 0.0026$$

This means that H_0 could have been rejected at significance level $\alpha = 0.0026$ which is much stronger than rejecting it at $\alpha = 0.05$.
Hence, the P-value gives more information than rejecting or not rejecting H_0 at a fixed α.



(A) Observed $\bar{X} = 101.5$

$z = 3$

very rare event

Could have moved critical value all the way here and still reject H_0

(B) Since this is a two-tailed test, the critical value on this side mirrors the right tail

(C) Therefore

$P = 2 \times 0.0013 =$

0.0026

Linear Regression

In engineering we often have more than one variable in an application and it is known that there exists some inherent relationship among the variables. We will study the case with 2 variables.

Example: Variable 1: Distance to transmitter : X

Variable 2: Wireless signal strength : Y

Lets assume a linear relationship between Y and x is reasonable:

$$Y = \alpha + \beta X$$

Dependent variable ↑ Y; ↑ α intercept; ↑ β slope; ↑ X Independent variable (Regressor)

Note: Don't confuse α, β with the type I & II error probabilities in hypothesis testing.

If the relationship between Y and X is exact, then it is a deterministic relationship.

In real applications, there are many sources of randomness:

- measurement noise
- the linear regression model might be an approximation to a much more complicated and possibly unknown relationship
- other factors (variables) not considered in model

Randomness means the same value of X does not always give the same value of Y (non-deterministic)

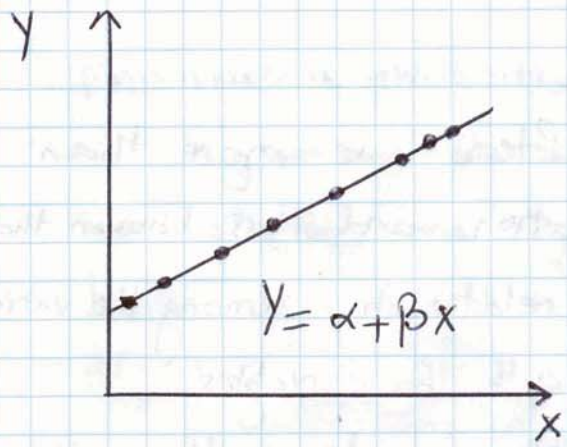
Simple Linear Regression Model:

$$Y = \alpha + \beta X + \varepsilon$$

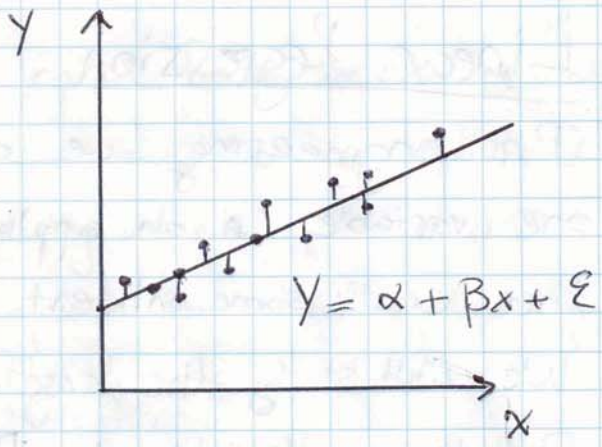
Dependent variable → Y; Unknown intercept → α; Unknown slope → β; Regressor → X; ε ← Random variable

$E[\varepsilon] = 0$

σ^2 variance



A deterministic relationship



A non-deterministic relationship

Important: ε is not a fixed number, it is a random variable which takes on a different value at each data point. The ε value at each point is shown as bars to the line $Y = \alpha + \beta x$ in the plot on the right above.

The only conditions on ε are $E[\varepsilon] = 0$

$Var[\varepsilon] = \sigma^2$

* $E[\varepsilon]$ implies that at a specific x, the Y values are distributed around the true (population) regression line.
* In practice α and β are unknown and must be estimated from data.

More advanced topics:

a) If there are more than one regressor variables

$$Y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$$

Example: Electromagnetic absorption in head due to cell phone use: Y

Signal frequency: x_1

Signal strength: x_2

Head size: x_3

b) If the relationship is nonlinear. There are many non-linear models, but a simple one would look like:

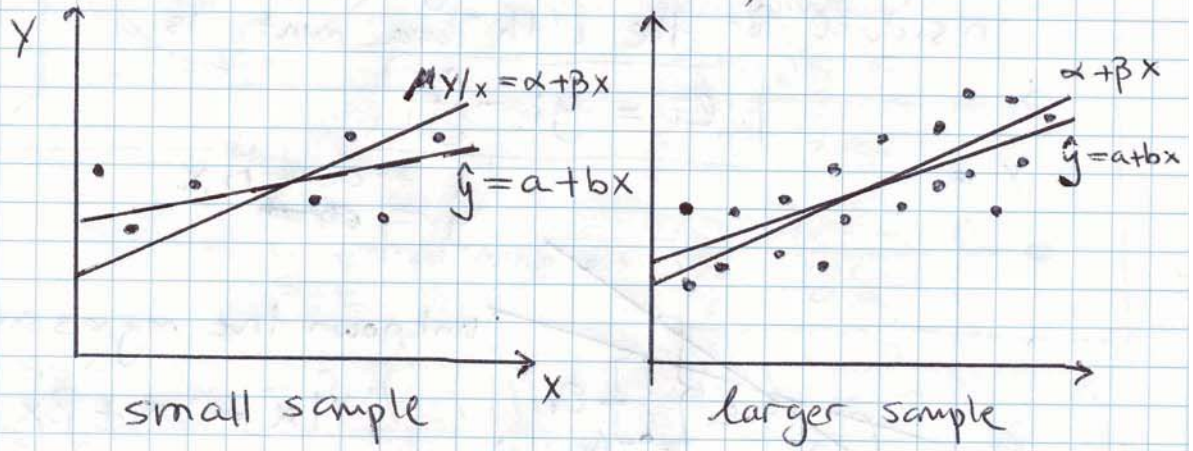
$$Y = \alpha + \beta x + \gamma x^2 + \varepsilon$$

This is a quadratic relationship.

Finding the parameters α and β from data

Example: In the example with Y signal strength and X distance to transmitter, we are given the pairs of observations (sample) (x_i, y_i) $i = 1$ to n and we want to estimate α, β.

Let a be our estimate of the true population parameter α. Let b be our estimate of the true population parameter β. Just like the sample mean $\bar{x}$, a and b depend on the particular sample (they are random!) The larger the sample size the closer a should be to α and b to β (just like $\bar{x}$ to μ)



$\mu_{Y|x} = \alpha + \beta x$: expected value of Y given X (true regression)
$Y = \alpha + \beta X$

$\hat{y} = a + bx$: fitted regression

$\hat{y}$ ↑ predicted (fitted) value

Conceptual model errors ε_i

$$Y = \alpha + \beta X + \varepsilon$$

$$E[Y] = E[\underbrace{\alpha + \beta X}_{\text{these are not random}} + \varepsilon]$$

$$= \alpha + \beta X + E[\varepsilon]$$

$E[\varepsilon] = 0$ assumption about ε

$$\mu_{Y|X} = E[Y] = \alpha + \beta X$$

Let x_i be the data points for x

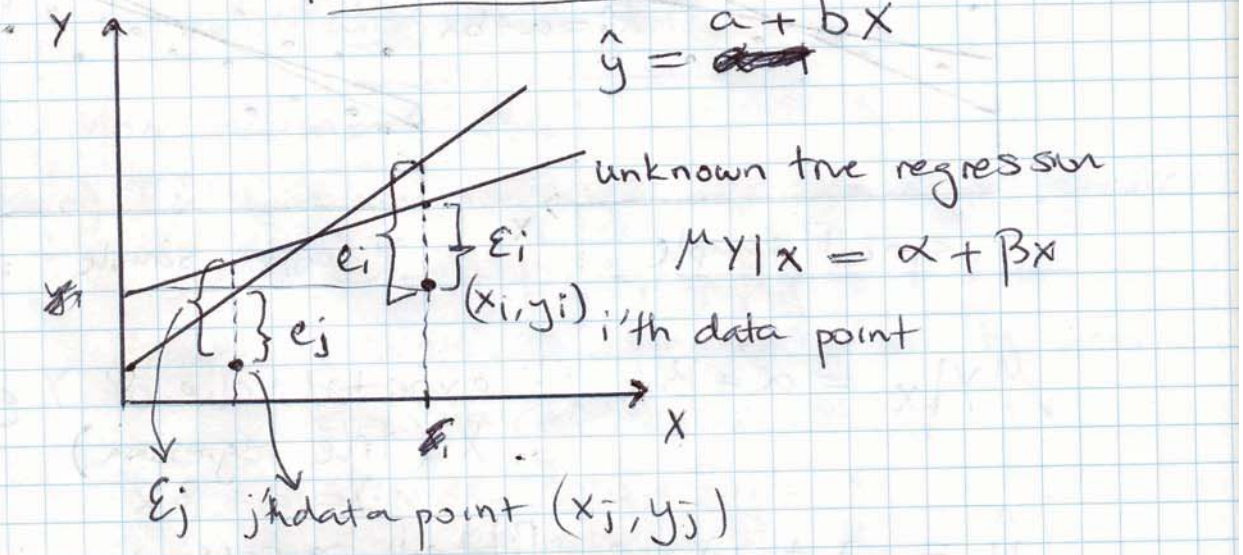
$$y_i = \alpha + \beta x_i + \varepsilon_i$$

$$\varepsilon_i = y_i - (\alpha + \beta x_i)$$ are the conceptual errors (realizations of the random var ε)

Residuals: error in fit

Given a set of data (sample) (x_i, y_i) $i = 1$ to n and a fitted model $\hat{y}_i = a + b x_i$ then the residual for the i'th data point is

$$\boxed{e_i = y_i - \hat{y}_i}$$



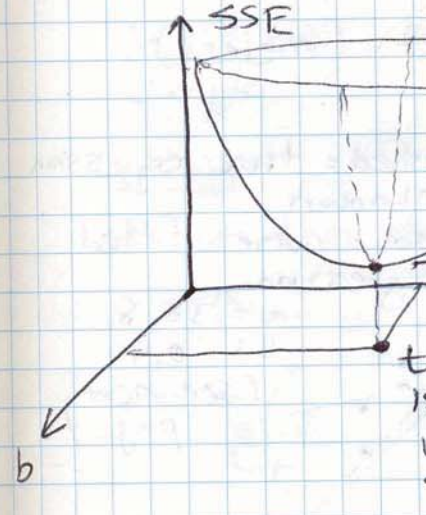
Least squares fitting method

One way to choose reasonable values for a and b is to choose them to minimize the sum of squared residuals.

$$SSE(\text{sum of squared errors}) = \sum_{i=1}^{n} e_i^2 \quad \leftarrow \text{Note this is } e_i \text{ not } \varepsilon_i$$

$$= \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - a - bx_i)^2$$

To minimize SSE with respect to a and b, from calculus we know that the partial derivatives of SSE with respect to a and b must be 0.



$$SSE = \sum_{i=1}^{n} (y_i - a - x_i)^2$$

at this point: *

$$\textcircled{1} \quad \frac{\partial SSE}{\partial a} = 0$$

$$\textcircled{2} \quad \frac{\partial SSE}{\partial b} = 0$$

$$\frac{\partial SSE}{\partial a} = -2 \sum_{i=1}^{n} (y_i - a - bx_i) = 0$$

$$\frac{\partial SSE}{\partial b} = -2 \sum_{i=1}^{n} (y_i - a - bx_i) x_i = 0$$

Rearranging terms gives:

$$\textcircled{1} \quad na + b \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i$$

$$\textcircled{2} \quad a \sum_{i=1}^{n} x_i + b \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i y_i$$

Solving ① and ② simultaneously gives the following formulas for a and b:

$$b = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

and $a = \bar{y} - b\bar{x}$

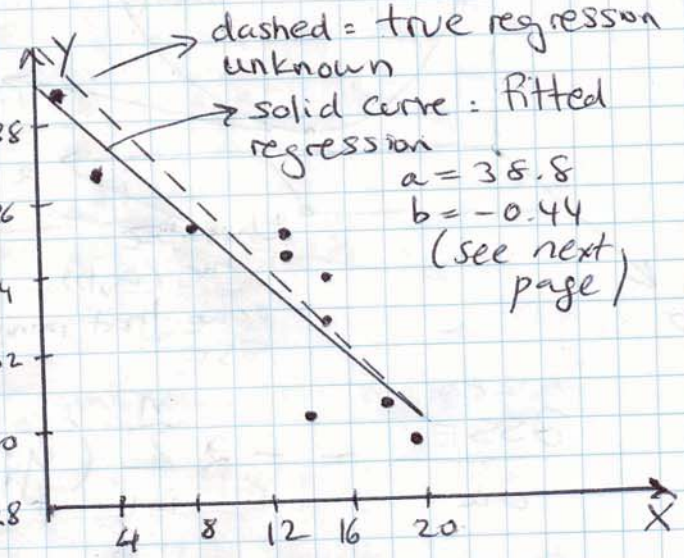
where $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$, $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$

Example Y: signal strength (dB)

X: distance to transmitter (meters)

10 measurements

i	x_i	y_i
1	13	34.4
2	1	38.4
3	17	30.4
4	19	29.7
5	14	30.1
6	15	33.9
7	15	32.8
8	8	35.2
9	13	34.9
10	3	36.8



Notice that for the same x value we have two different y values (non-deterministic)
Same for $x = 13$

Least squares fitting

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = 11.8 \qquad \bar{y} = 33.66$$

$$b = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sum_{i=1}^{n}(x_i-\bar{x})^2} = \frac{-137.58}{315.6} = -0.4359 \approx -0.44$$

$$a = \bar{y} - b\bar{x} = 38.8 \Rightarrow \boxed{\hat{y} = 38.8 - 0.44x}$$

Normally, we wouldn't know the true values α, β, but in this case, I generated the data myself according to $Y = 40 - 0.5x + \varepsilon$ where ε was normally distributed with mean 0 and variance 1. So α was -0.5 and β was 40. Our estimates are quite close, but they could be better with a larger sample.

Question: Predict what the signal strength would be if the distance was 10 meters and 26 meters?

$$\hat{y} = a + bx = 38.8 - 0.44x$$

so for $x = 10$ $\quad \hat{y} = 34.4$ dB

for $x = 26$ $\quad \hat{y} = 27.36$ dB

Note: Sometimes we have control over for which x_i we make measurements. For instance, we could have measured signal strength y_i for x_i at regularly spaced intervals

2, 4, 6, 8, 10, 12, 14, 16, 18, 20 meters.