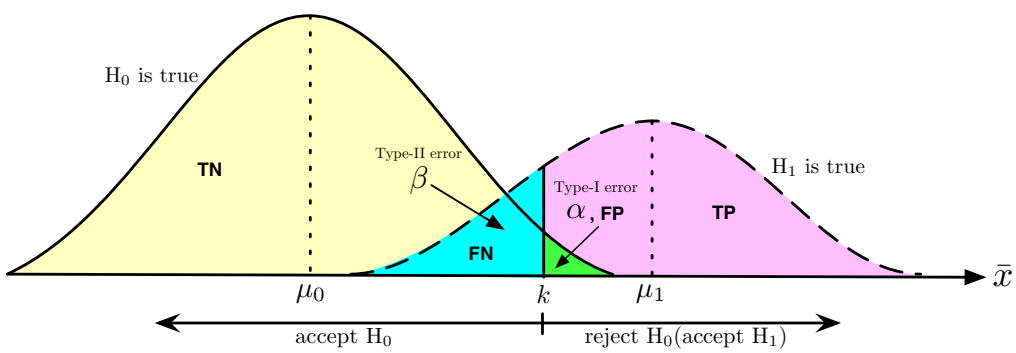
Handout 2

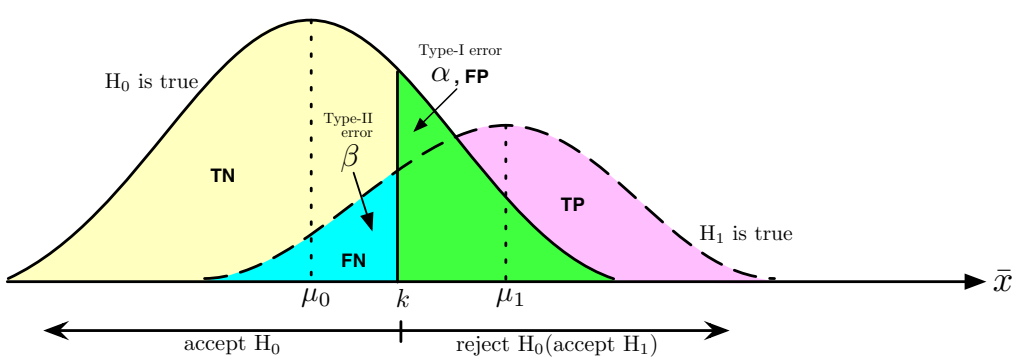
1 Confusion Matrix (Test Outcomes, 0: negative, 1: positive)

	Predict 0 (negative, rejection) (consider H_0 ture, fail to reject H_0)	Predict 1 (positive, identification) (consider H_0 false, reject H_0)
True 0 $Y = 0$ (negative) (H_0 is true)	**True Negative, TN** Specificity, Confidence level: $1 - \alpha$ *correct rejection* $\Longrightarrow$ find good person innocent [Correct Decision, $\mathrm{P} = 1 - \alpha$]	**False Positive, FP** Type-I error (α error), Significance level: α *false alarm, overestimation, incorrect identification* $\Longrightarrow$ find good person guilty [Wrong Decision: reject a true H_0, $\mathrm{P} = \alpha$]
True 1 $Y = 1$ (positive) (H_1 is true)	**False Negative, FN** Type-II error (β error), 1-Power: β *miss, underestimation, incorrect rejection* $\Longrightarrow$ find bad person innocent [Wrong Decision: accept a false H_0, $\mathrm{P} = \beta$]	**True Positive, TP** Sensitivity, Power: $1 - \beta$ *hit, correct identification* $\Longrightarrow$ find bad person guilty [Correct Decision, $\mathrm{P} = 1 - \beta$]

- α (significance level) $= P$(make a Type-I error) = **size** of a test; $\beta = P$(make a Type-II error)
- Power (the probability when H_1 is true and you successfully reject H_0) = $1 - \beta$
- Larger sample size: $\alpha, \beta \downarrow$ ($n \uparrow \Longrightarrow$ greater power/sensitivity)

2 Graphical Representation (If $\alpha \downarrow$, then $\beta \uparrow$, critical value: k)





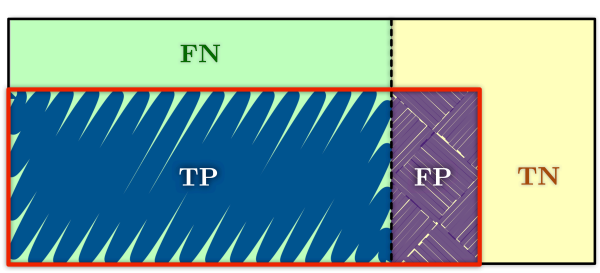
3 Type-I Error (α) and Type-II Error (β)

- The risk of committing a Type-I error is the significance level (α) chosen for the study (default: 5%).
- 5% significance level: Our results only have 5% (or less) chance of occurring if H_0 (no effect) is true.
  $\Longrightarrow$ There's a 5% risk of concluding that a difference exists when there is no actual difference.
  $\Longrightarrow$ If data show sufficient evidence supporting our result, then instead of obtaining rare effect ($\leq 5\%$), it must be that our premise (assuming H_0 is correct) is wrong $\Longrightarrow$ Reject H_0, conclude effect exists.
- A Type-II error (β) means failing to conclude that there was an effect when there actually was. This may happen when our study lacks statistical power to detect an effect of a certain size.
- Power ($1-\beta$) is whether a test can correctly detect a real effect when there is one (default: 80%, $\beta = 0.2$). The higher the power, the lower the probability of making a Type-II error.
- Power is determined by
  1. Size of the effect: Larger effects are more easily detected (require less power).
  2. Measurement error: Systematic and random errors in data reduce power.
  3. Sample size: Larger samples reduce sampling error and increase power.
  4. Significance level: Increasing the significance level increases power ($\alpha \uparrow \Longrightarrow \beta \downarrow \Longrightarrow$ power $\uparrow$).

4 Discussions

- Consider a criminal justice example. Innocent until proven guilty $\Longrightarrow$ H_0 : not guilty, H_1 : guilty.
- Y (outcome, verdict) is either 0 (not guilty) or 1 (guilty): $Y \in \{0, 1\}$.
- A *Type-I error* is made when you find a person guilty but he is innocent (wrongly convicted).
- A *Type-II error* is made when a person is found innocent but is guilty.
- When conducting a *trial*, making Type-I error (miscarriage of justice) is more serious than Type-II error.
- For a test of *censoring cancer*, you don't want to make Type-II error. For misdiagnosis, instead of telling a sick patient ($Y = 1$) that he does not have cancer and delays radiation therapy, you should make a healthy person ($Y = 0$) go through additional tests.
- If you are writing a *mailbox spam filter*, then you don't want to make Type-I error. Letting spam ($Y = 1$) slip into the inbox is better than accidentally throwing an important email ($Y = 0$) into the junk box.

5 Classification Context



The red boxed area is predicted **Positive**. The green area is positive ($Y = 1$) and the yellow area is negative ($Y = 0$). For instance, the purple area is negative but predicted positive which is *false positive* (FP).

5.1 Precision: the correct positive in your prediction

Precision is also referred to the *positive predictive value* which is defined as

$$\text{Precision} = \frac{\text{the correct "positive" you predict}}{\text{all "positive" you predict (red box)}} = \frac{\text{TP}}{\text{TP} + \text{FP}}. \tag{1}$$

This is the proportion of the data points the model says was relevant (positive) actually were relevant.
Model focus fixing FP→ $\alpha \downarrow \rightarrow$ precision ↑ (recall ↓). Not affected by data imbalance in rare disease study (negative $\gg$ positive cases).

5.2 Recall ($1 - \beta$): the true positive covered by your prediction

Recall is also referred to the *true positive rate* or *sensitivity (power)* which is defined as

$$\text{Recall} = \frac{\text{the correct "positive" you predict}}{\text{all true "positive" (green area)}} = \frac{\text{TP}}{\text{TP} + \text{FN}}. \tag{2}$$

This is the ability to find all relevant instances in the data. Model focus fixing FN→ $\beta \downarrow \rightarrow$ recall ↑ (percision ↓).

5.3 F1 Score (Evaluate model performance)

F1 Score is the harmonic mean of Precision and Recall ($\because$ trade-off between recall and precision)

$$F_1 = \frac{2}{\frac{1}{\text{Recall}} + \frac{1}{\text{Precision}}} = \frac{2(\text{Recall} \times \text{Precision})}{\text{Recall} + \text{Precision}} = \frac{2\text{TP}}{2\text{TP} + \text{FN} + \text{FP}}. \tag{3}$$

The harmonic mean punishes extreme values. A model with a precision of 1 and recall of 0 has an average of 0.5 but an F_1 of 0.

5.4 Specificity ($1 - \alpha$): the true negative covered by your prediction

Specificity (confidence level) is also referred to *true negative rate* (or *1 - false positive rate*) which is defined as

$$\text{Specificity} = \frac{\text{the correct "negative" you predict}}{\text{all true "negative" (yellow area)}} = \frac{\text{TN}}{\text{TN} + \text{FP}}. \tag{4}$$

5.5 False Positive Rate

This is the probability that a false alarm will be raised and that a positive result will be given when the true value is negative which is defined as

$$\text{False Positive Rate} = \frac{\text{FP}}{\text{TN} + \text{FP}} = 1 - \text{Specificity (true negative rate)}. \tag{5}$$

5.6 Accuracy

This is the probability that measures overall, **how often is our model correct**?

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{TN} + \text{FN}}. \tag{6}$$

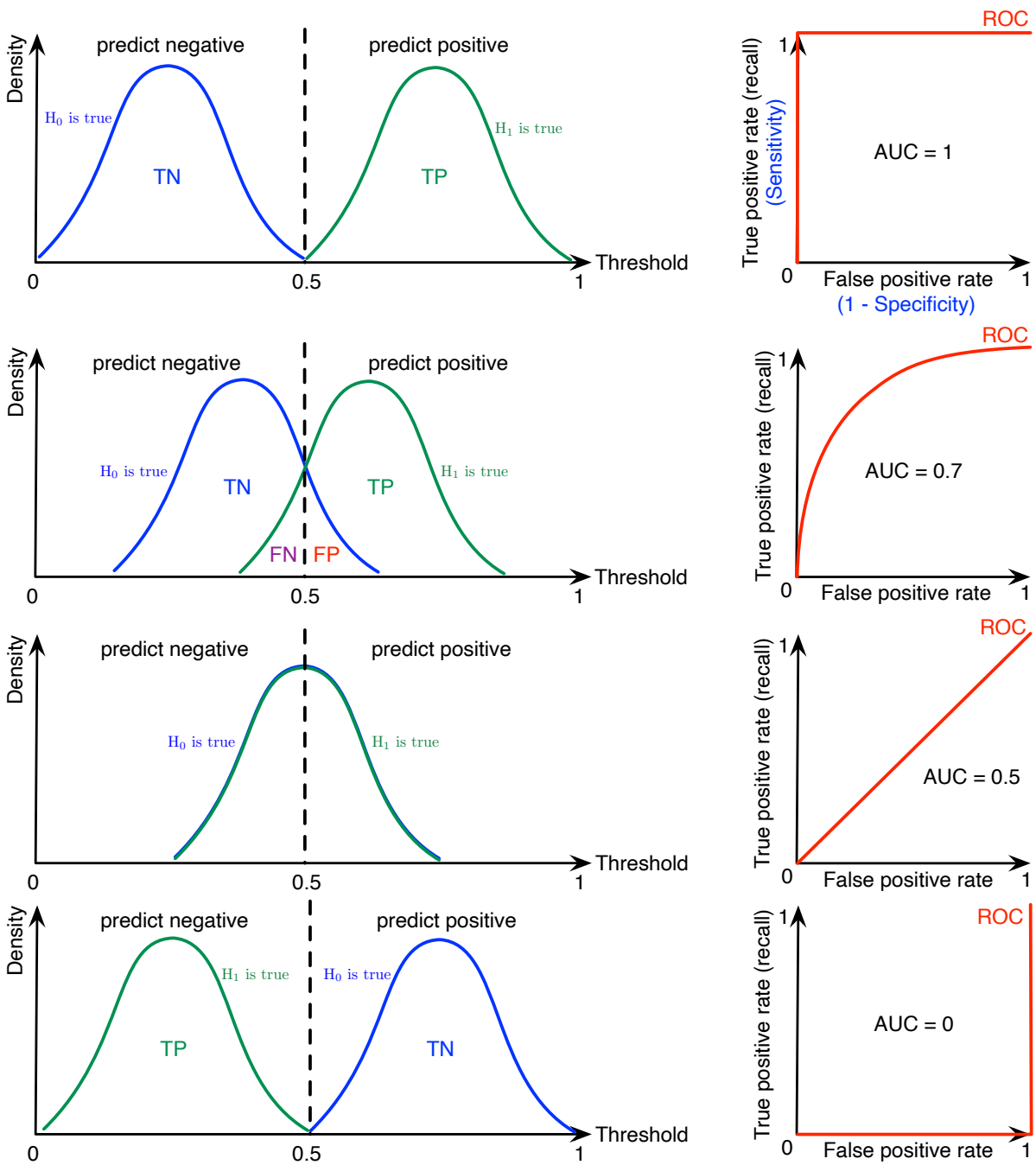
- Accuracy can tell us immediately whether a model is being trained correctly and how it may perform generally.
- Accuracy does not do well (not a good metric) when you have a severe **class imbalance**.
- For example, in email spam detection:

		actual	
		Not Spam	Spam
predicted	Not Spam	98 (TN)	0 (FN)
	Spam	1 (FP)	1 (TP)

  - The accuracy is $\frac{1+98}{1+1+98+0} = 99\%$.
  - However, if we look at the two predicted spam emails, only one of them is actual spam $\Longrightarrow$ 50%. Indicating that this model with an 99% accuracy does well in filtering non-spam emails. However, it has only 50% chances to correctly filter out spam emails.

6 Performance Measurement

- ROC (line): Receiver operating characteristic curve (summarizes **all** the confusion matrices that each classification threshold produced → Identifies the best threshold for making a decision).
  - To compute the points on ROC, we could evaluate a logistic regression many times with different classification thresholds.
  - Often replace $false\ positiv\ rate$ with $precision$ (proportion of positives that correctly classified).
- AUC (area): Area Under the ROC curve (Aggregate measure of model performance. Makes it easy to compare one ROC curve to another → Help decide which categorization method is better).
  - For binary classification: ROC-AUC = $\dfrac{\text{recall (}or\text{ sensitivity) + specificity (}or\text{ 1 - false positive rate)}}{2}$



7 Statistical significance vs Power

We will discuss the relationship between significance and power. If the test for an effect is statistically significant, do we still need to care about the power of the test?

7.1 Statistical significance

- P-value is the probability of observing the data (or an effect that is more extreme) if H_0 is true.
  - A small P-value ($P < 0.05$) is evidence against $H_0 \rightarrow$ reject H_0 and accept H_1.
- Significant P-value of a test: avoids *Type-I error*.
  - **Type-I error**: Rejecting H_0 when H_0 is true; **inferring an effect when none exists**.

7.2 Power

- The power of a test is the probability that the test will correctly reject H_0 when H_1 is true.
  - The probability of detecting an effect if there really is an effect.
- Power of a test: avoids *Type-II error*.
  - **Type-II error**: Failing to reject H_0 when H_1 is true; **failing to infer an effect when one exists**.

7.3 Power vs Effect Size

- If the actual effect (e.g. difference between groups, correlation strength) in the population is small, the test might not have enough power to **consistently** detect it as statistically significant.
- Large effects are easier to detect than small effects.

7.4 Power vs Sample Size

- When the sample size increases, the power of the test also increases, making it more likely to detect an effect if it's truly there.
- A study with a small sample size may have a significant p-value by chance, but the power might be low (high chance of committing a Type II error if the effect size is not large).
  - The (significant) effect might be very small and potentially not of practical importance.
- A very large sample size can detect even trivial differences as statistically significant.

A study with low power might not be considered robust even if it yields a statistically significant P-value. Since There's a good chance that the study might not find a significant result if it were **repeated**.

When designing a study, researchers often conduct a **power analysis** to determine the required **sample size** to detect a given **effect size** with a specified **level of confidence**.