# SESSION 10: PROBABILITY THEORY (VI)

Dr Gang Li

Deakin University, Geelong, Australia

2018-11-06

## Table of Content





# Statistical Inference



## Probability vs Statistics

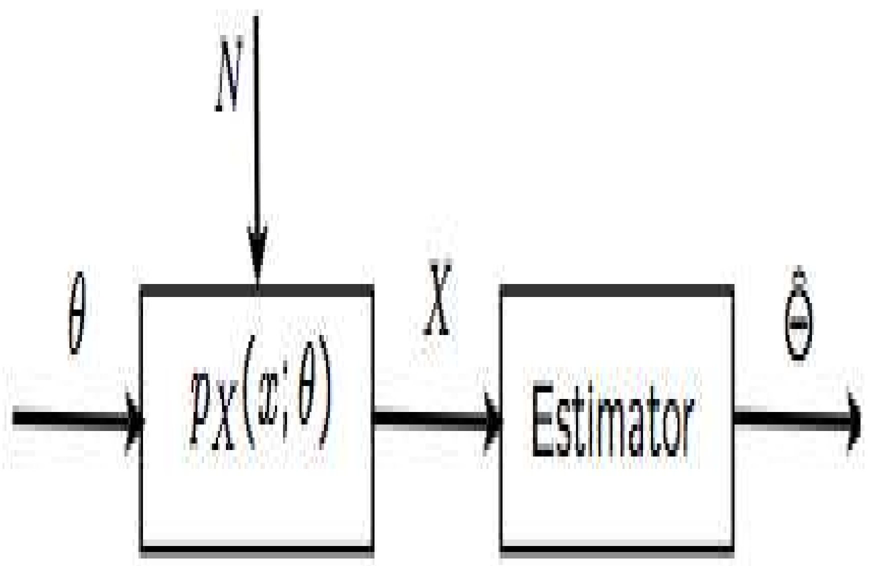
**Probability**

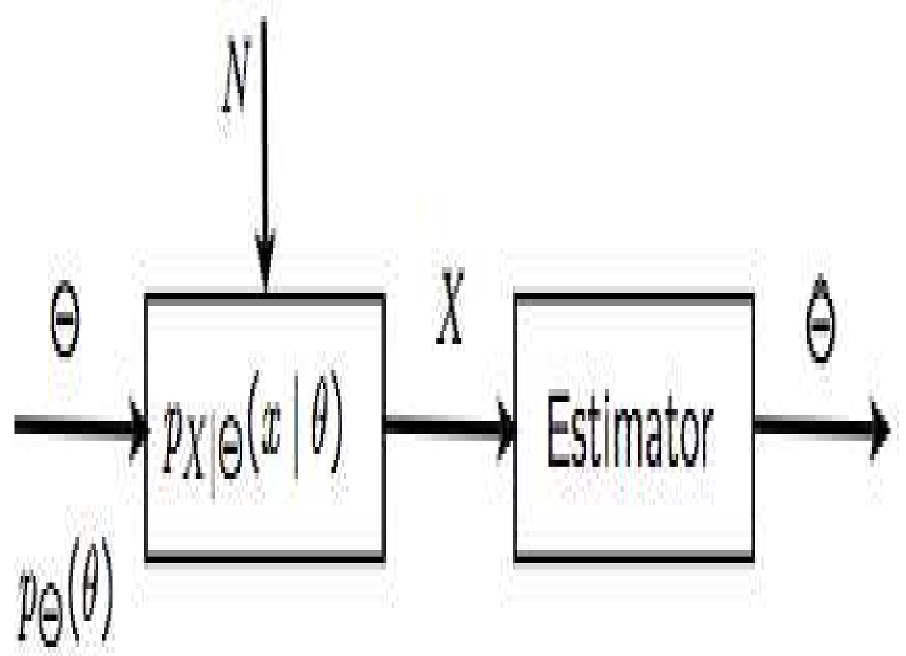
- We assume a fully specified probabilistic model that obeys the axioms.
- We then use mathematical methods to quantify the consequences of this model, or answer various questions of interest.
- Every unambiguous question has a unique correct answer, though this answer is sometimes hard to find.

**Statistics**

- It involves an element of art
- Several reasonable methods may exist, yielding different answers
- No principled way for selecting the 'best' method, unless one makes several assumptions and imposes additional constraints on the inference problem

### Frequentist vs Bayesian Statistics




**Frequentist Statistics**

- $\theta$ is treated as deterministic quantities, that happen to be unknown
- it strives to develop an estimate of $\theta$ that as some performance guarantees.
- we are not dealing with a single probabilistic model, but rather with multiple candidate probabilistic models, one for each possible value of $\theta$.

**Bayesian Statistics**

- It views the model as chosen randomly from a given model class.
- $\theta$ is treated as a random variable that characterizes the model, and by postulating a *prior* probability distribution $p_\Theta(\theta)$.
- Use priors and Bayes rule to derive a *posterior* probability distribution $p_{\Theta|X}(\theta|x)$, which captures all the information that $x$ can provide about $\theta$.



## Bayesian Statistics 6 / 24

### Bayesian Statistics

**Bayesian Statistics** treats unknown parameters as random variables with known prior distributions.

**Parameter Estimation** generates estimates that are close to the true values of the parameters in some probabilistic sense.

**Hypothesis Testing** the unknown parameter takes one of the finite number of values, corresponding to competing hypotheses; We want to choose one to achieve a small probability of error.

**Bayesian Inference Methods**

**MAP Rule** out of the possible parameter values/hypotheses, select one with maximum conditional or posterior probability given the data.

**Least Mean Squares (LMS)** Select an estimator/function of the data that minimizes the mean squared error between the parameter and its estimate.

**Linear Least Mean Squares (LMS)** Select an estimator/function which is a linear function of the data and minimizes the mean squared error between the parameter and its estimate.

## MAP Rule

**Defn** Given the value $x$ of the observation, we select a value of $\theta$, denoted by $\hat{\theta}$, that maximizes the posterior distribution $p_{\Theta|X}(\theta|x)$, or $f_{\Theta|X}(\theta|x)$ if $\Theta$ is continuous. This is the *Maximum a Posteriori Probability* (MAP) rule.

**Priors** Bayesian methods provide a way to include prior information in a systematic way: $p(\theta)$.

**Non-informative Prior** represent lack of information, but it has one major flaw: if it is flat in one parameterization it will not be flat in most other parameterizations.
https://normaldeviate.wordpress.com/2012/12/08/flat-priors-in-flatland-stones-paradox/

**Single Answer** If interested in a single answer, though single answers can be misleading!

**MAP**

$$p_{\Theta|X}(\theta^*|x) = \max_{\theta} p_{\Theta|X}(\theta|x)$$

which minimizes the probability of error, often used in hypothesis testing

**Conditional Expectation**

$$E[\Theta|X=y] = \int \theta f_{\Theta|X}(\theta|x) d\theta$$



## Least Mean Squares (LMS) Estimation

**LMS** LMS estimates $\Theta$ with $\hat{\theta}$ so that the estimation error $E[(\Theta-\hat{\theta})^2]$ is least.

**In the Absence of Information** Estimating $\Theta$ with a constant $\hat{\theta}$, in the absence of an observation $X$.

- The estimation error $\hat{\theta}-\Theta$ is random, because $\Theta$ is random.
- but the *mean squared error* $E[(\Theta-\hat{\theta})^2]$ is a number that depends on $\hat{\theta}$, and can be minimized over $\hat{\theta}$.
- For any estimate $\hat{\theta}$, we have

$$E[(\Theta-\hat{\theta})^2] = var(\Theta-\hat{\theta}) + (E[\Theta-\hat{\theta}])^2 = var(\Theta) + (E[\Theta-\hat{\theta}])^2$$

  - The first one from $E(Z^2) = var(Z) + (E(Z)^2$
  - The second one from the $\hat{\theta}$ is a constant
  - We choose $\hat{\theta}$ to minimize $(E[\Theta-\hat{\theta}])^2$, which leads to $\hat{\theta} = E[\Theta]$.

**In the Observation of $X = x$** Estimating $\Theta$ with a constant $\hat{\theta}$, in the observation $X$.

- It is a new universe condition on $X = x$
- So the conditional expectation $E[\Theta|X = x$ minimizes the conditional mean squared error $E[(\Theta-\hat{\theta})^2|X = x]$ over all constants $\hat{\theta}$.
- $E(\Theta|X]$ minimizes $E[(\Theta - g(X))^2]$ over all estimators $g(\cdot)$

**Properties of LMS estimation** Estimator: $\hat{\Theta} = E[\Theta|X]$ Estimation error: $\tilde{\Theta} = \hat{\Theta} - \Theta$

- $E(\tilde{\Theta}) = 0$, and $E(\tilde{\Theta}|X = x) = E(\hat{\Theta} - \Theta|X = x) = E(\hat{\Theta}|X) - E(\Theta|X = x) = \hat{\Theta} - \hat{\Theta} = 0$, So $\hat{\Theta}$ is unbiased.
- $E(\tilde{\Theta}h(x)|x) = h(x)E(\tilde{\Theta}|x) = 0$ From the law of iterative expectations, we have $E(\tilde{\Theta}h(x) = 0$
- $Cov(\tilde{\Theta}h(x)) = E(\hat{\Theta}h(x)) - E(\hat{\Theta})E(h(x)) = 0$, So $Cov(\tilde{\Theta}\hat{\Theta}) = 0$.
- Since $\Theta = \hat{\Theta} - \tilde{\Theta}$, and their covariance is zero, we have $var(\Theta) = var(\hat{\Theta}) + var(\tilde{\Theta})$.

### Linear Least Mean Squares Estimation

**Defn** A linear estimator of a random variable $\Theta$, based on observations $X_1, \cdots, X_n$ has the form

$$\hat{\Theta} = \alpha_1 X_1 + \cdots + \alpha_n X_n + \beta$$

Given a particular choice of the scalars $\alpha_1, \cdots, \alpha_n, \beta$, the corresponding mean squared error is $E[(\Theta - \alpha_1 X_1 - \cdots - \alpha_n X_n - \beta)^2]$

**Best linear estimator**

$$\hat{\Theta}_L = E(\Theta) + \frac{Cov(X,\Theta)}{var(X)}(X - E[X])$$

- $\alpha = \frac{Cov(X,\Theta)}{var(X)} = \rho\frac{\sigma_\Theta}{\sigma_X}$, where $\rho = \frac{Cov(\Theta,X)}{\sigma_\Theta \sigma_X}$.
- With the MSE as $E[(\hat{\Theta} - \Theta)^2] = (1 - \rho^2)\sigma_\Theta^2$
- The formula only involves the means, variances, and the covariance of $\Theta$ and $X$.



## Classical Statistics



### Classical Statistics

**Classical Statistics** treats unknown parameters as constants to be determined. A separate probabilistic model is assumed for each possible value of the unknown parameter.

**Parameter Estimation** generates estimates that are nearly correct under any possible value of the unknown parameter.

**Hypothesis Testing** the unknown parameter takes finite number $m \geq 2$ of values, corresponding to competing hypotheses; We want to choose one to achieve a small probability of error under any of the possible hypotheses.

**Classical Inference Methods**

**MLE** Select the parameter that makes the observed data “most likely”, i.e., maximizes the probability of obtaining the data at hand.

**Linear Regression** Find the linear relation that matches best a set of data pairs, in the sense that it minimizes the sum of the squares of the discrepancies between the model and the data.

## ML Estimation

**Defn** Let the vector of observations $X = (X_1, \cdots, X_n)$ be described by $p_X(x;\theta)$ whose form depends on an unknown parameter $\theta$. Suppose we observe a particular value $x = (x_1, \cdots, x_n)$ of $X$, then the *Maximum Likelihood* (ML) estimation is a value of the parameter that maximizes the *likelihood function* $p_X(x_1, \cdots, x_n;\theta)$ over all $\theta$:

$$\hat{\theta}_n = \arg\max_{\theta} p_X(x_1, \cdots, x_n;\theta)$$

**Example** Suppose $X = (X_1, \cdots, X_n)$ are i.i.d. from $exp(\theta)$: $\theta e^{-\theta x}$

- $max_\theta \prod_{i=1}^n \theta e^{-\theta x}$
- Take the logarithm $max_\theta (n \log\theta - \theta \sum_{i=1}^n x_i)$
- $\hat{\theta}_{ML} = \frac{n}{x_1 + \cdots + x_n}$

**Desirable Properties** Let $\hat{\Theta}_n$ be an estimator of an unknown parameter $\theta$, that is, a function of $n$ observations $X_1, \cdots, X_n$ whose distribution depends on $\theta$.

**Estimation Error** denoted by $\tilde{\Theta}_n = \hat{\Theta}_n - \theta$

**Bias** of the estimator $\hat{\Theta}_n$, denoted by $b_\theta(\hat{\Theta}) = E_\theta[\hat{\Theta}_n] - \theta$

**Unbiased** If $E_\theta[\hat{\Theta}_n] = \theta$, for every possible value of $\theta$

**Asymptotically unbiased** if $\lim_{n\to\infty} E_\theta[\hat{\Theta}_n] = \theta$, for every possible value of $\theta$

**Consistent** if the sequence $\hat{\Theta}_n$ converges to the true value of $\theta$, in probability, for every possible value of $\theta$

**MSE (Bias Variance Decomposition)** $E[(\hat{\Theta}_n - \theta)^2] = var(\hat{\Theta}_n - \theta) + (E[\hat{\Theta}_n - \theta])^2 + \sigma_\epsilon^2 = var(\hat{\Theta}_n) + (bias)^2 + \sigma_\epsilon^2$

**Example** Suppose $X = (X_1, \cdots, X_n)$ are i.i.d. mean $\theta$ variance $\sigma^2$

$$X_i = \theta + W_i$$

with $W_i$ i.i.d. mean 0, variance $\sigma^2$.

- We have the sample mean $\hat{\Theta}_n = M_n = \frac{X_1 + \cdots + X_n}{n}$
  - It is unbiased $E(\tilde{\Theta}) = 0$
  - From WLLN: $\hat{\Theta}_n \to \theta$, so it is consistent.
  - MSE: $E[(\hat{\Theta}_n - \theta)^2 = \frac{\sigma^2}{n}]$



## Bias-Variance Tradeoff

**Defn** There are three kinds of errors in the estimation:

**The error due to bias** is taken as the difference between the expected prediction of our model and the correct value which we are trying to predict

**The error due to variance** is taken as the variability of a model prediction for a given data point

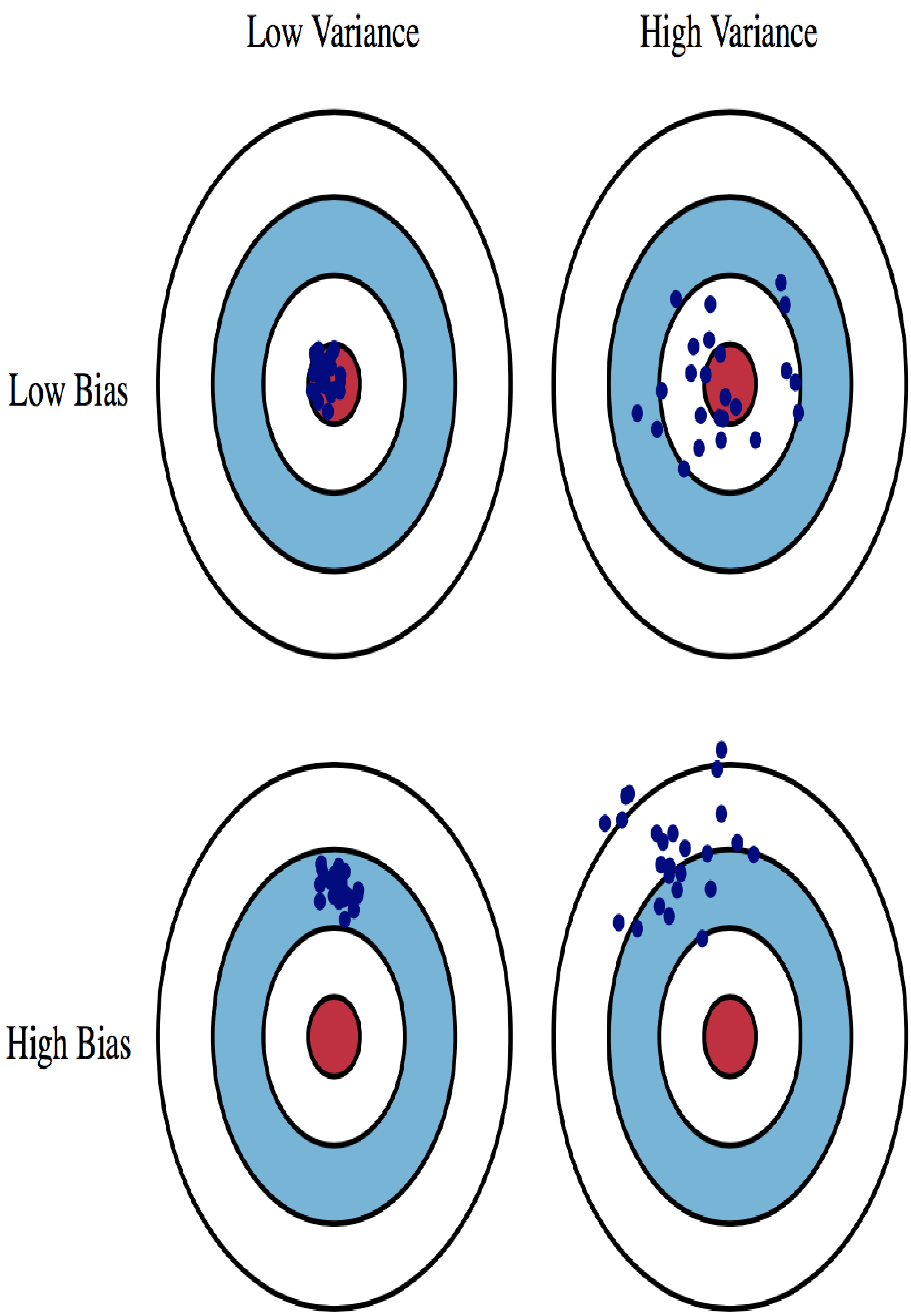
**The irreducible error** is the noise term in the true relationship that cannot fundamentally be reduced by any model

*Proof.*

$$\begin{aligned} E[(\hat{\Theta}_n - \theta)^2] &= var(\hat{\Theta}_n - \theta) + (E[\hat{\Theta}_n - \theta])^2 + \sigma_\epsilon^2 \\ &= var(\hat{\Theta}_n) + (bias)^2 + \sigma_\epsilon^2 \\ &= Variance + Bias^2 + IrreducibleError \end{aligned}$$

□

**Bias-Variance Tradeoff**




## Confidence Interval

**Defn** Let us fix a desired *confidence level*, $1-\alpha$, where $\alpha$ is typically a small number. We then replace the point estimate $\hat{\Theta}_n$ by a lower estimator $\hat{\Theta}_n^-$ and an upper estimator $\hat{\Theta}_n^+$, so that $P_\theta(\hat{\Theta}_n^- \leq \theta \leq \hat{\Theta}_n^+) \geq 1-\alpha$ for every possible value of $\theta$. Here both $\hat{\Theta}_n^-$ and $\hat{\Theta}_n^+$ are functions of observations, and hence random variables whose distributions depend on $\theta$. We call $[\hat{\Theta}_n^-, \hat{\Theta}_n^+]$ a $1-\alpha$ *confidence interval*.

**Example** Suppose $X = (X_1, \cdots, X_n)$ are i.i.d., CI in estimation of the mean $\hat{\Theta}_n = \frac{X_1 + \cdots + X_n}{n}$

- From normal table $\Phi(1.96) = 1 - 0.05/2$
- From CLT, we have $P(\frac{|\hat{\Theta}_n - \theta|}{\sigma/\sqrt{n}} \leq 1.96) \approx 0.95$
- Then we have $P(\hat{\Theta}_n - \frac{1.96\sigma}{\sqrt{n}} \leq \theta \leq \hat{\Theta}_n + \frac{1.96\sigma}{\sqrt{n}}) \approx 0.95$
- More generally, let $z$ be s.t. $\Phi(z) = 1 - \alpha/2$ [a], then $P(\hat{\Theta}_n - \frac{z\sigma}{\sqrt{n}} \leq \theta \leq \hat{\Theta}_n + \frac{z\sigma}{\sqrt{n}}) \approx 1 - \alpha$

**Unknown $\sigma$** In the case of unknown $\sigma$,

**Option** 1 use the upper bound on $\sigma$, especially if $X_i$ Bernoulli, we have $\sigma \leq 1/2$.
**Option** 2 use ad hoc estimate of $\sigma$, if $X_i$ Bernoulli, we have $\sigma = \sqrt{\hat{\Theta}(1-\hat{\Theta})}$
**Option** 3 use generic estimate of the variance.

- Start from $\sigma^2 = E[(X_i - \theta)^2]$, $\hat{\sigma}_n^2 = \frac{1}{n}\sum_{i=1}^n (X_i - \theta)^2 \to \sigma^2$, but we don't know $\theta$.
- $\hat{S}_n^2 = \frac{1}{n-1}\sum_{i=1}^n (X_i - \hat{\Theta}_n)^2 \to \sigma^2$, unbiased: $E[\hat{S}_n^2] = \sigma^2$

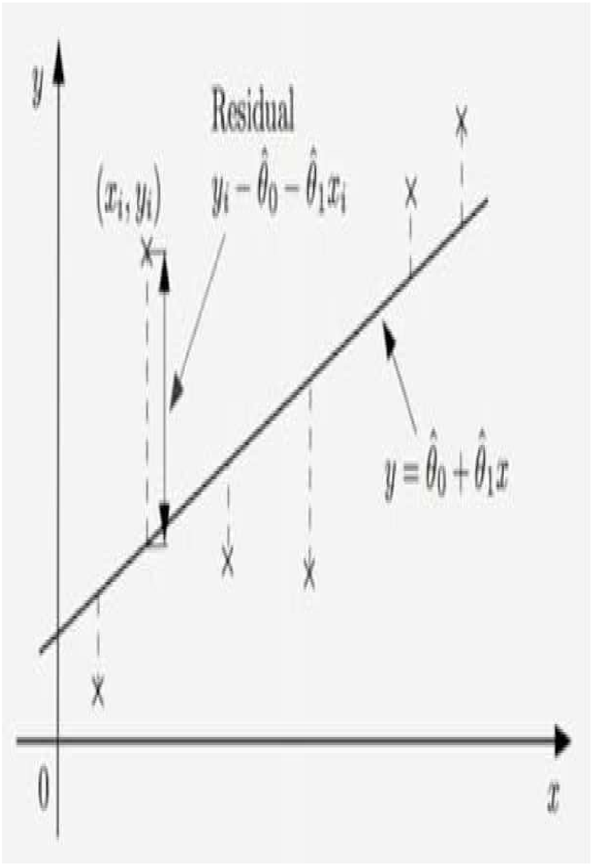
[a] When $n$ is small, $\hat{\Theta}_n^2$ is only an approximation to the true variance, and the random variable $T_n = \frac{\sqrt{n}(\hat{\Theta}_n - \theta)}{\hat{\Theta}_n}$ is not normal, but the *t-distribution* with $n-1$ degrees of freedom. $\bar{}_{n-1}(z) = 1 - \alpha/2$, where $\bar{}_{n-1}(z)$ is the CDF of the *t-distribution* with $n-1$ degrees of freedom. Check http://www.sumsar.net/blog/2013/12/t-as-a-mixture-of-normals/.

## Linear Regression

Defn

We wish to model the relation between $x$ and $y$, based on data set $(x_i, y_i)$, $i = 1, \cdots, n$. Assume a linear model of the form $y \approx \theta_0 + \theta_1 x$, where $\theta_0$ and $\theta_1$ are unknown parameters, the objective is to solve:

$$\min_{\theta_0, \theta_1} \sum_{i=1}^{n} (y_i - \theta_0 - \theta_1 x_i)^2$$




**One Interpretation**

$Y_i = \theta_0 + \theta_1 X_i + W_i$, with $W_i \sim N(0, \sigma^2)$

- Likelihood function is $c \cdot \exp\{-\frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - \theta_0 - \theta_1 x_i)^2\}$.
- Take logs, same as the linear regression objective.
- Least squares $\leftrightarrow$ pretend $W_i$ is i.i.d. normal.

**Solution** $\bar{x} = \frac{x_1 + \cdots + x_n}{n}$, $\bar{y} = \frac{y_1 + \cdots + y_n}{n}$

- Assume $W$ is independent of $X$ and with zero mean
- $E[Y] = \theta_0 + \theta_1 E[X]$ so we have $\theta_0 = E[Y] - \theta_1 E[X]$, hence, $\hat{\theta}_0 = \bar{y} - \hat{\theta}_1 \bar{x}$, though $\hat{\theta}_1$ is unknown.
- Assume for simplicity $E[X] = E[W] = 0$, $YX = \theta_0 X + \theta_1 X^2 + XW$. Take expectation on both sides $Cov(X, Y) = 0 + \theta_1 Var(X) + 0$, hence, $\hat{\theta}_1 = \frac{Cov(X,Y)}{Var(X)}$.

**Multiple Linear Regression** $y \approx \theta_0 + \theta x + \theta' x' + \theta'' x''$, typically resort to linear algebra

**Standard Error** an estimate of $\sigma$

**Explanatory Power** $R^2 = \frac{Var(Y|X)}{var(Y)}$, a measure of explanatory power: when $R^2$ is less, it means whenever I know $X$, $Y$ is well known, or $X$ explains $1 - R^2$ percentage of $Y$.

**Common Pitfalls**

**Heteroskedasticity** when $var(W_i)$ is strongly affected by the value of $x_i$.

**Multicollinearity** when two indicator variables $x$ and $x'$ bear a strong relation.

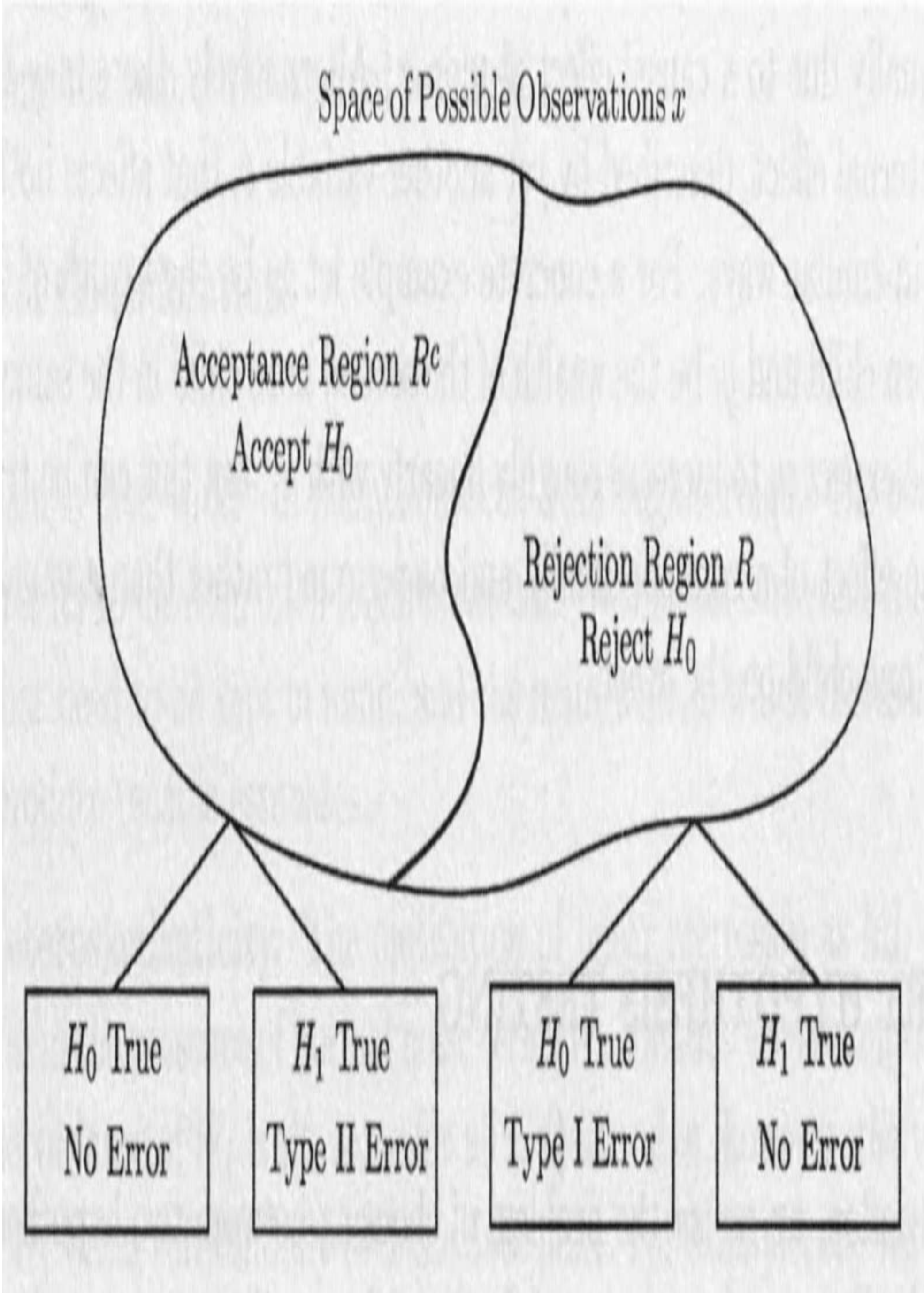
**Overfitting** The danger of producing a model that fits the data well, but is otherwise useless. A rule of thumb, there should be at least five or preferably ten times more data points than there are parameters to be estimated.

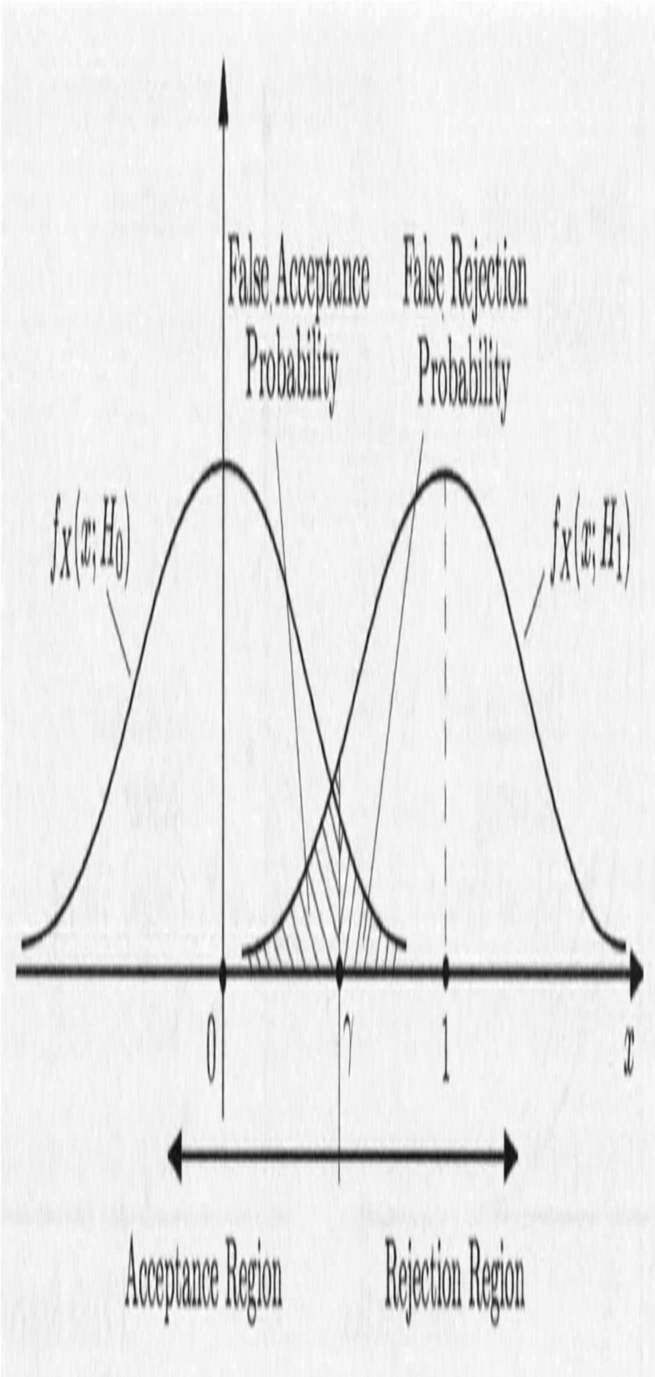
**Casuality** a linear relation should not be mistaken for a causal relation.

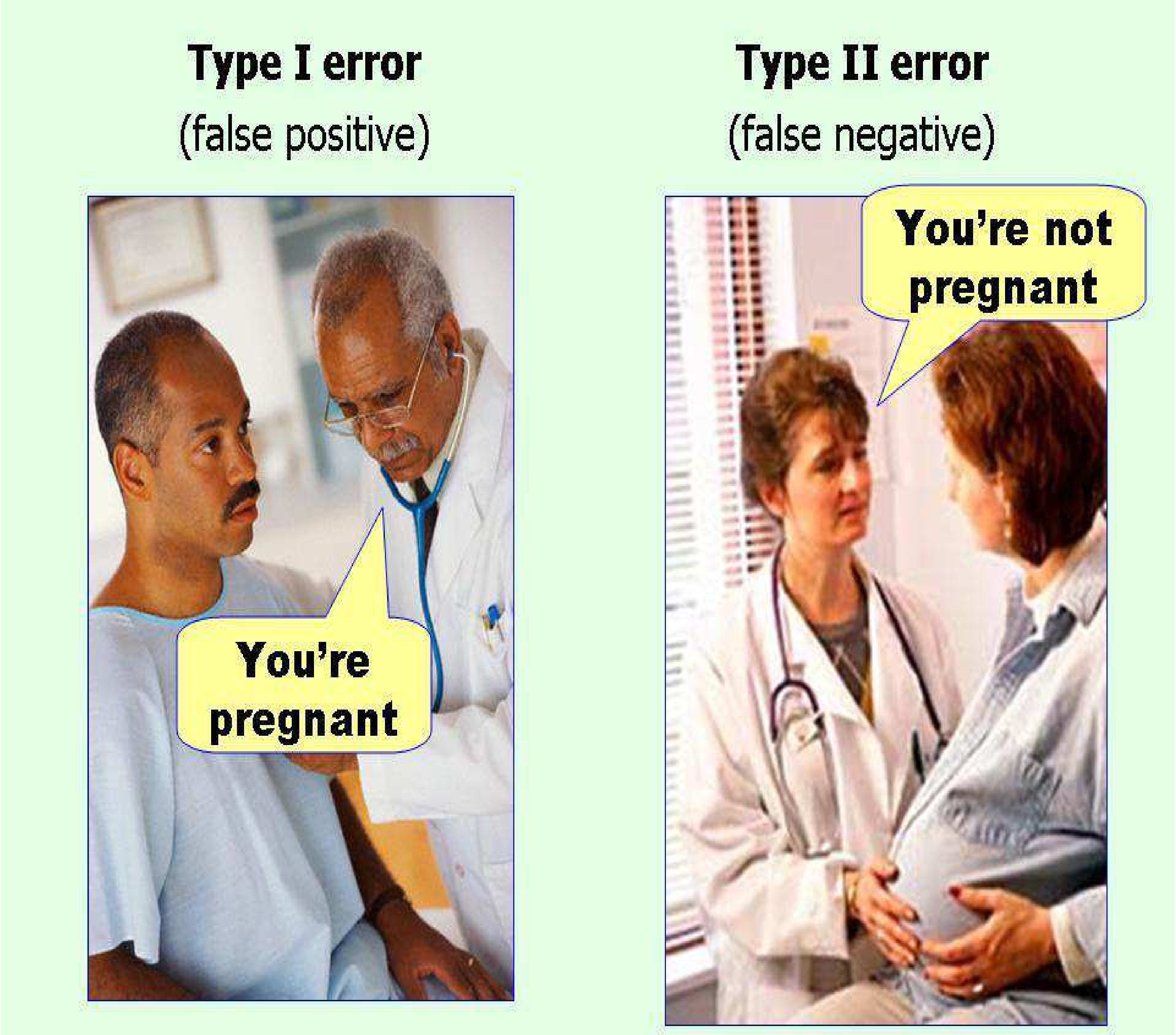
## Binary Hypothesis Testing

**Defn** Assume no prior probability, we choose between two hypotheses $H_0$ and $H_1$. Hypothesis $H_0$ is often called the *null hypothesis*, and $H_1$ the *alternative hypothesis*. This indicate that $H_0$ plays the role of a default model, to be proved or disproved on the basis of the available data.

***Type I* and *Type II* Error**



## Likelihood Ration Test (LRT)

Defn

Assume no prior probability, we choose between two hypotheses $H_0$ and $H_1$. Define the *likelihood ratio* by

$$L(x) = \frac{p_X(x; H_1)}{p_X(x; H_0)}$$

where $p_X(x; H)$ denotes the PMF or PDF of the vector $X$ under hypothesis $H$.

- Start with a target value $\alpha$ for the false rejection probability; typically 0.1, 0.05 or 0.01;
- Choose the *critical value* for $\xi$ such that the false rejection probability is equal to $\alpha$:

$$P(L(X) > \xi; H_0) = \alpha$$

- Once the value $x$ of $X$ is observed, reject $H_0$ if $L(X) > \xi$.



## Significant Testing

Defn

When composite hypotheses involved, namely, no two well-specified alternatives, we wish to determine on the basis of observations $X = (X_1, \cdots, X_n)$ whether the null hypothesis $H_0$ should be rejected or not.

*General Steps.*

- Choose a *statistic* $S$, namely a scalar random variable that will summarize the data to be obtained;
- Determine the *shape of the rejection region* by specifying the set of values of $S$ for which $H_0$ will be rejected as a function of a yet undetermined critical value $\xi$
- Choose the *significance level*, namely the desired probability $\alpha$ of a false rejection of $H_0$
- Choose the *critical value* $\xi$ so that the probability of false rejection is equal to $\alpha$. At this point, the rejection region is completely determined.

□

*Example.* Got $S = 472$ heads in $n = 1000$ tosses; is this coin fair?
$H_0$: $p = \frac{1}{2}$ versus $H_1$: $p \neq \frac{1}{2}$

- Choose a *statistic* $S$
- Determine the *rejection region*, $|S - \frac{n}{2}| > \xi$
- Choose the *significance level* $\alpha = 0.05$
- Choose the *critical value* $\xi$ so that

$$P(reject H_0; H_0) = \alpha$$

  Using the CLT, we have $\xi = 31$:

$$P(|S - 500| \leq 31; H_0) \approx 0.95$$

- As $|S - 500| = 28 < \xi$, so $H_0$ is not rejected at the 5% level.

□

## NHST and P-Value

Defn

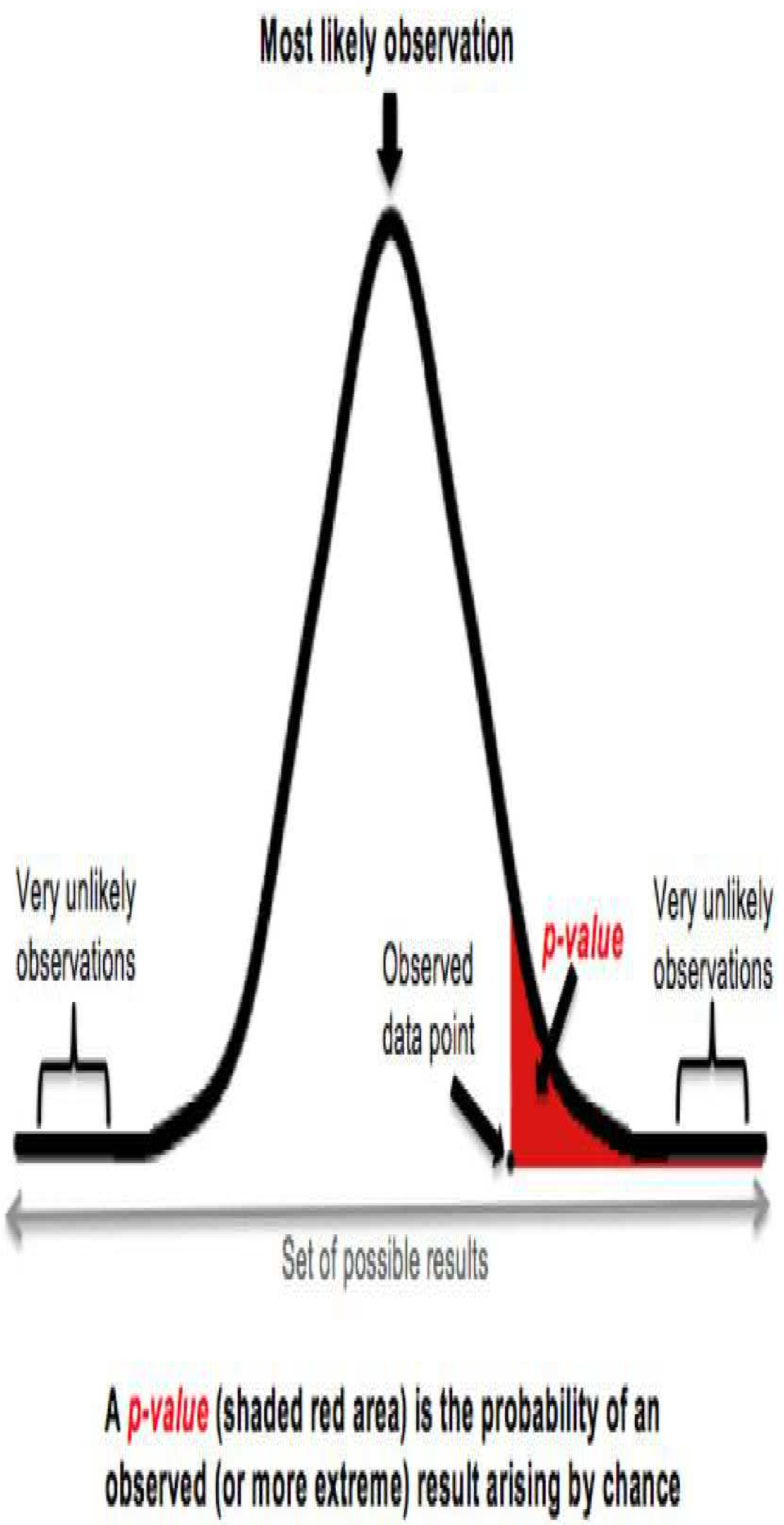
Warning: *P-values, the 'gold standard' of statistical validity, are not as reliable as many scientists assume.*

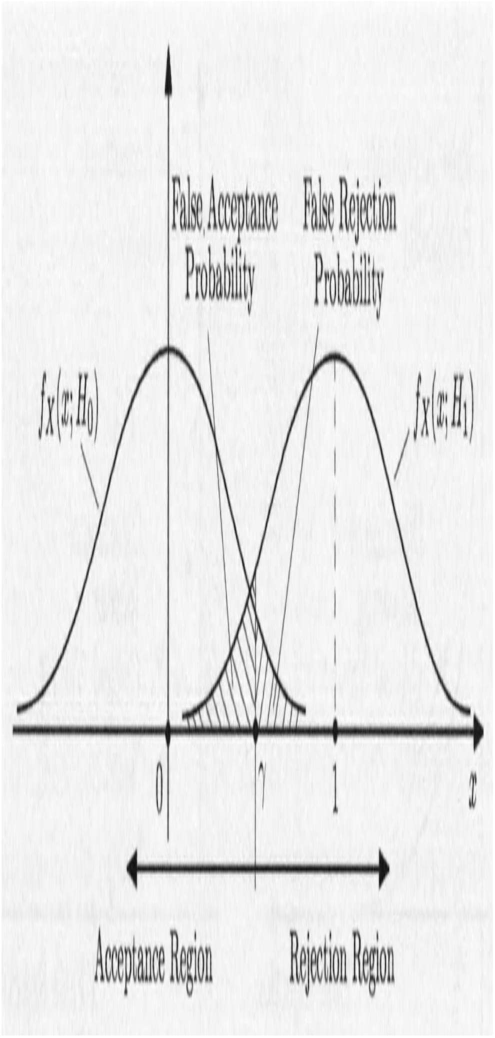
**Null Hypothesis Significance Testing** (NHST)

- State a null hypothesis: that is, there is no effect.
- Calculate the p value, which is the probability of getting results like ours - if the null hypothesis is true.
- If p is sufficiently small, reject the null hypothesis and sound the trumpets: our effect is not zero, it's statistically significant!

Regina Nuzzo. Statistical errors. *Nature*, 506(Feburary):150–152, 2014
John P. A. Ioannidis. Why most published research findings are false. *PLoS Medicine*, 2(8):696–701, 2005
American Statistical Association. ASA statement on statistical significance and p-values. *The American Statistician*, 70(2):129–133, 2016

**NHST and P-Value**




Defn

Consider a $2 \times 2$ table in which research findings are compared against the gold standard of true relationships.

| Findings | True (Y) | True (No) | Total |
|---|---|---|---|
| Find (Y) | $c(1-\beta)R/(R+1)$ | $c\alpha/(R+1)$ | $c(R+\alpha-\beta R)/(R+1)$ |
| Find (N) | $c\beta R/(R+1)$ | $c(1-\alpha)/(R+1)$ | $c(1-\alpha+\beta R)/(R+1)$ |
| Total | $cR/(R+1)$ | $c/(R+1)$ | $c$ |

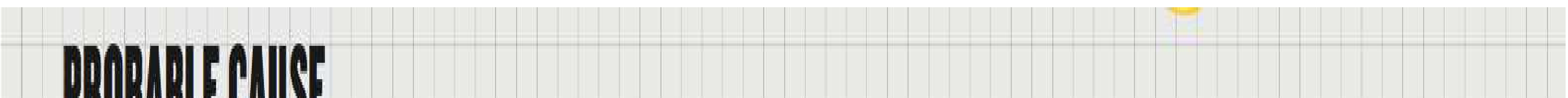
- Assume either there is only one true relationship (among many hypothesized) or the power is similar to find any of the several existing true relationships.
- Let $R$ be the ratio of the number of "true relationships" to "no relationships" among those tested in the field. $R = \frac{P_Y}{P_N}$
- The pre-study probability of a relationship being true is $P_Y = \frac{R}{R+1}$
- The probability of a study finding a true relationship reflects the power $1-\beta$ (one minus the Type II error rate)
- The probability of claiming a relationship when none truly exists reflects the Type I error rate, $\alpha$.

When the finding shows $Yes$, how likely the truth is really $Yes$?

- The post study probability that is true is the positive predictive value (PPV), which is

$$PPV = P(Truth = Yes|Finding = Yes) = \frac{(1-\beta)R}{(R-\beta R+\alpha)}$$

  - When $(1-\beta)R > \alpha$, we have $PPV > 50\%$, namely it is more likely true than false.
  - If we take p value 0.05, namely here $\alpha = 0.05$, this means that $PPV$ will be likely true than false when $(1-\beta)R > 0.05$.
  - If just report 0.05, it actually does not imply anything on the findings.



## Questions?



## Contact Information

Associate Professor **Gang Li**
School of Information Technology
Deakin University, Australia

GANGLI@TULIP.ORG.AU
TEAM FOR UNIVERSAL LEARNING AND INTELLIGENT PROCESSING