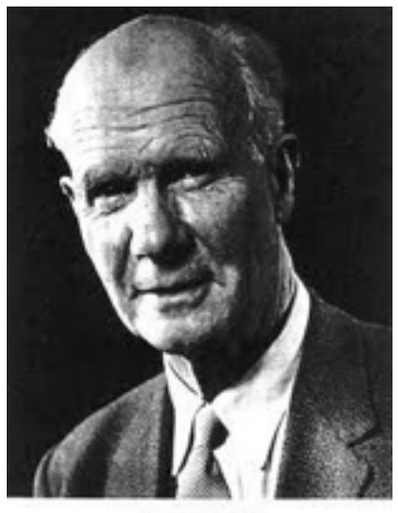
Statistics for Phys 494:

Hypothesis testing

Paul A. Wiggins

R. A. Fisher

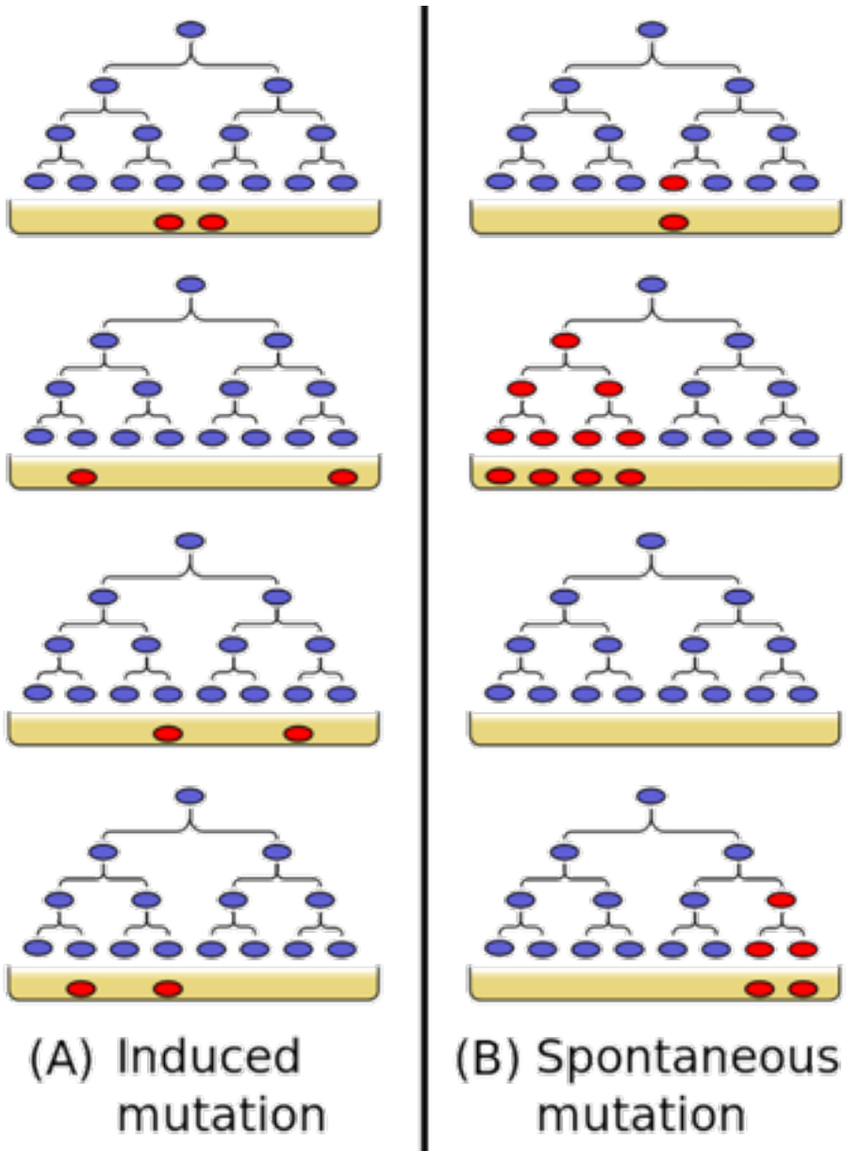
J. Neyman



E. Pearson

Phys: “We understand fluctuations!”
Biologists: Luria and Delbrück

- Lamarkian versus Darwinian evolution:



- Do mutations arise **spontaneously** or are they **induced** by stress
- Lamark: I**nduced**
- Darwin: **Spontaneous**
- Experiment to determine mechanism:
  - propagate cells
  - stress
  - measure distribution of mutants
  - “Jackpot” versus “Poisson” distribution

… other disciplines often have a better understanding of statistics that physics students…

- **Critical reading** of scientific papers demands an understanding of basic **hypothesis testing**!
    - What is a P-value?
- Questions posed in one's **own research**:
    - How do I know if my model is "right"?
    - Can I calculate the probability that the model is "right"?
    - How do I measure statistical evidence?

Summary of hypothesis testing

1 Null and Alternative Hypotheses

2 Test Statistic

3 *P*-Value

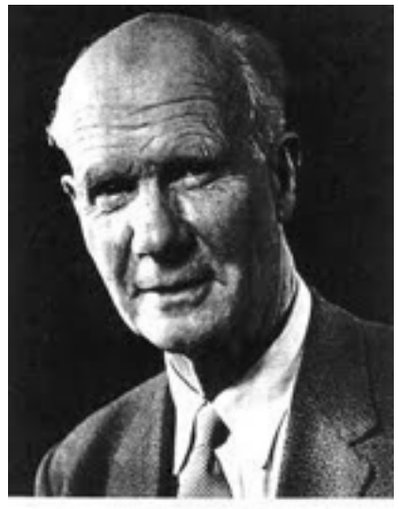
4 Significance Level

5 Power

6 Examples

R. A. Fisher

J. Neyman

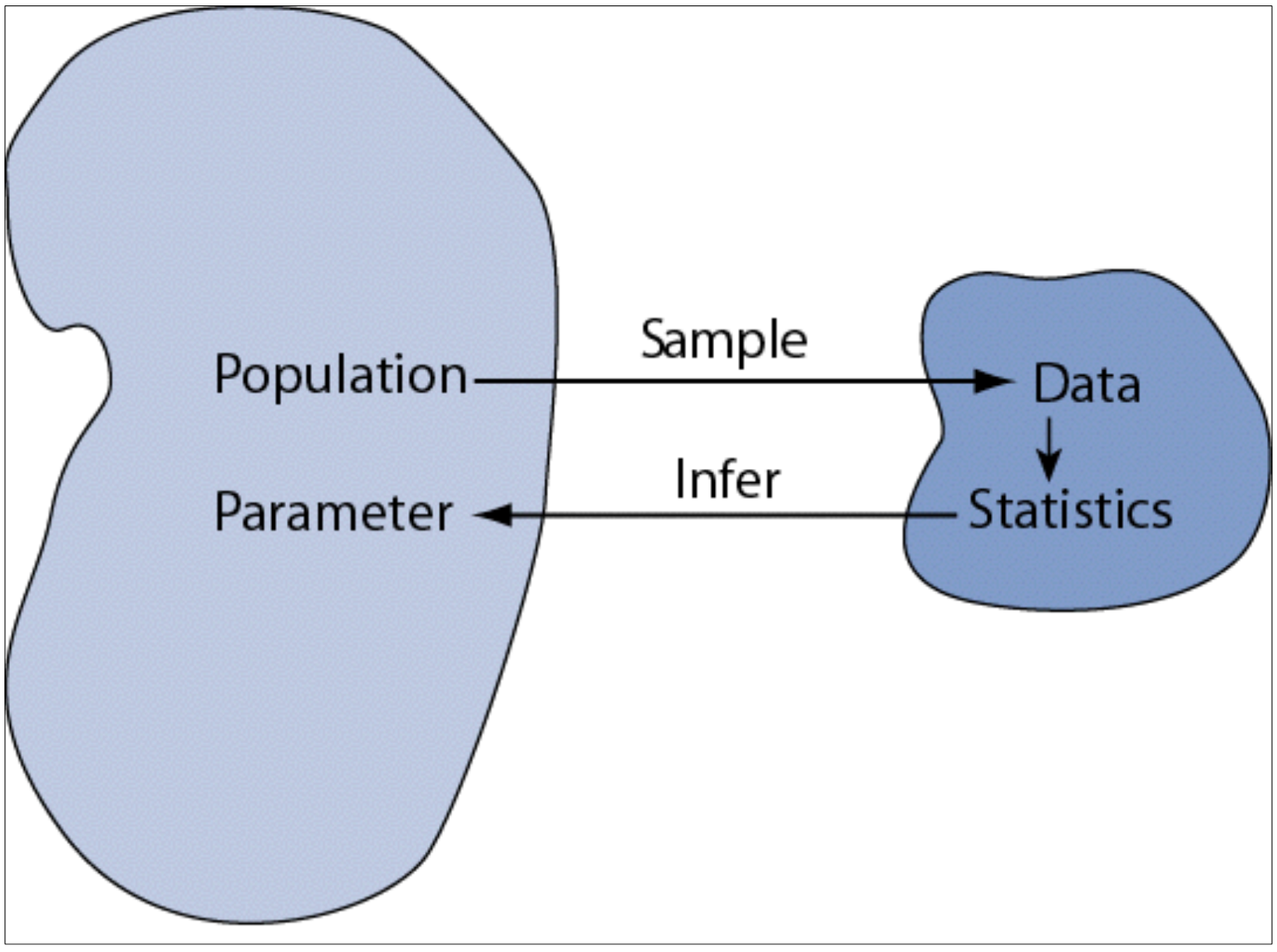


E. Pearson

Basic statistical vocabulary:

- **Observations/Data** ≡ X_i
- **Population** ≡ True distribution of X values
- **Sample** ≡ $X^N \equiv \{X_i\}_{i=1..N}$
- **Statistical inference** ≡ generalizing from a sample to a population with calculated degree of certainty
- Two related types of inference
  - **Hypothesis testing**
  - **Estimation**
- **Hypothesis** ≡ MODEL of the data
- **Parameter** ≡ variable in a MODEL,
  e.g. population mean *μ*
- **Statistic** ≡ function of observed sample,
  e.g. empirical mean ($\overline{X}$)

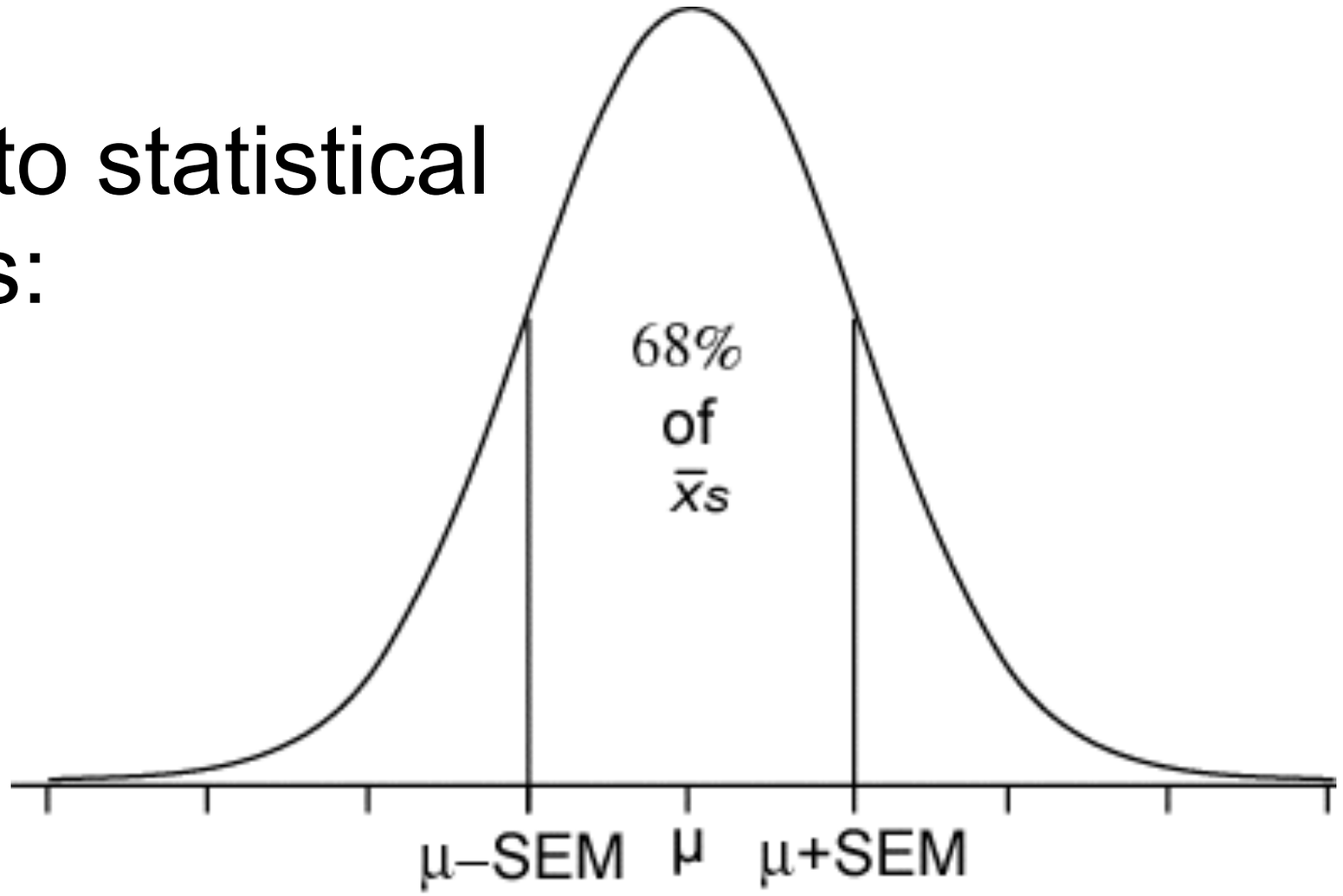
What is estimation in inference?



Statistics have **statistical variation**

- The sample/empirical mean (a statistic) **estimates** the true mean (a parameter)
- … subject to statistical fluctuations:

$$\sigma_{\overline{x}} \equiv \frac{\sigma}{\sqrt{N}}$$



$\sigma_{\overline{x}} \equiv$ Standard Error of the Mean (SEM)

Hypothesis Testing

- … also called ***significance testing***

- Measure how consistent the data is with a MODEL (i.e. hypothesis)

- 5 steps

- *Each* element of the procedure must be understood

Null hypothesis Test steps

A. Define **null** and **alternative** hypotheses

B. Choose & compute a **test statistic**

C. Choose a **significance** (α)**/confidence level**

D. Compute the P-value

E. Interpret data by comparing p and α
(**Reject** or **Accept** the **null hypothesis**)

Null and Alternative Hypotheses

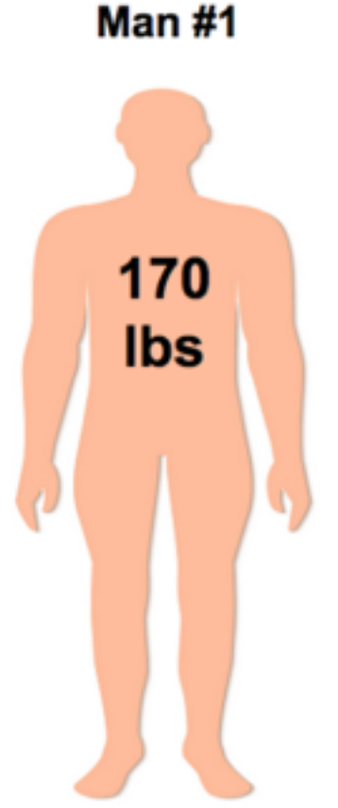
- Convert the research question to null and alternative hypotheses
  - The **null hypothesis (H_0)** – "No new physics"
  - The **alternative hypothesis (H_1)** – "New physics"
- Collect data and seek evidence to **reject** H_0 (and therefore support H_1)

Illustrative example

- **The problem:** In 1970, men had a mean (μ) body weight of 170 lb with a standard deviation (σ) of 40 lb. Are men heavier today?

- **Null hypothesis:**
  H_0: μ = 170 lb (“no difference”)

- The **alternative hypothesis:**
  H_1: μ > 170 lb



Test Statistic

- How do you choose a test statistic?
  - **Literature** defines good statistics for different types of problems
    - In W*eight Problem* we use the z-score:

$$z(\overline{x}) = \frac{\overline{x} - \mu}{\sigma_{\overline{x}}}$$

- General approach: **Likelihood ratio test**

Illustrative Example: *z* statistic

- For the illustrative example, $\mu_0 = 170$
- We know $\sigma = 40$
- collect a sample of $N = 64$. Therefore

$$\sigma_{\overline{x}} = \sigma/\sqrt{N} = 5 \text{ lb}$$

- If we found a sample mean of 173 lb, then

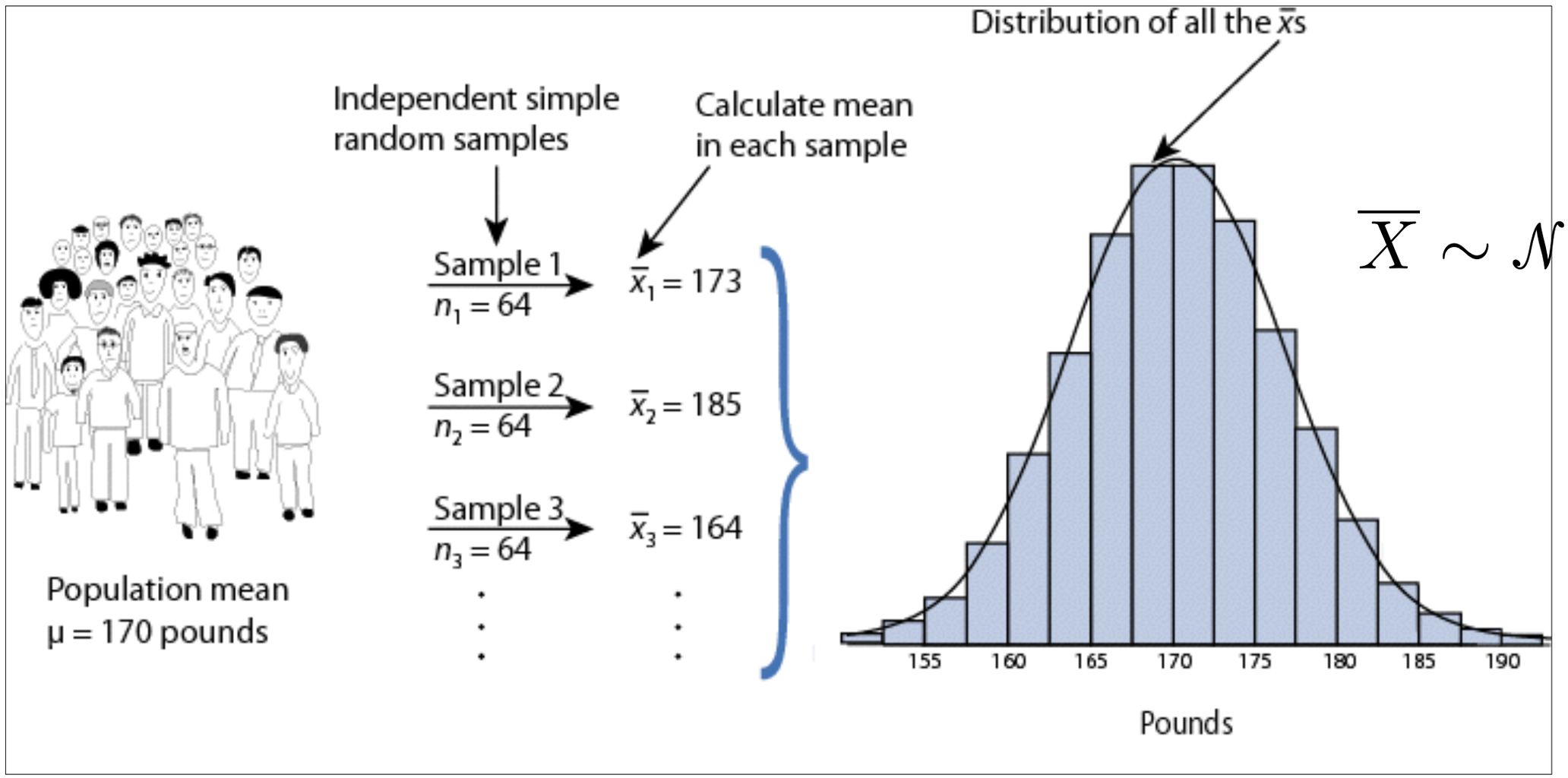
$$z = \frac{\overline{x} - \mu}{\sigma_{\overline{x}}} = 0.60$$

- If we found a sample mean of 185 lb, then

$$z = \frac{\overline{x} - \mu}{\sigma_{\overline{x}}} = 3.0$$

Reasoning behind choosing *z* *as the test statistic*

z is a measure of the distance from the expected value divided by the expected variation

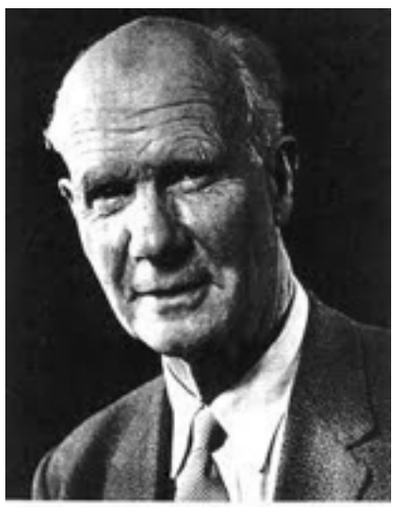


$$\overline{X} \sim \mathcal{N}(170, 5)$$

Significance/Confidence Level

- How SURE do we need to be?
- **Significance**: α ≡ probability of erroneously rejecting H_0
- **Confidence**: γ ≡ (1-α) probability of correctly accepting H_0
- Set α threshold (e.g., let α = .10, .05, *or whatever*)
- N&P: Choose AHEAD of time (stops fine tuning)

J. Neyman



E. Pearson

The *P*-value

- The *P*-value answer the question: What is the probability of the observed test statistic or one more extreme **when H_0 *is true*?**
- Treat the test statistics as a random variable:

$$Z \sim z(X^N) \qquad \text{where} \qquad X^N \sim H_0$$

- Compute the Cumulative Probability Distribution (CDF) of the test statistic assuming the null hypothesis:

$$F_Z(z) = \int_{-\infty}^{z} dz' \; f_Z(z')$$

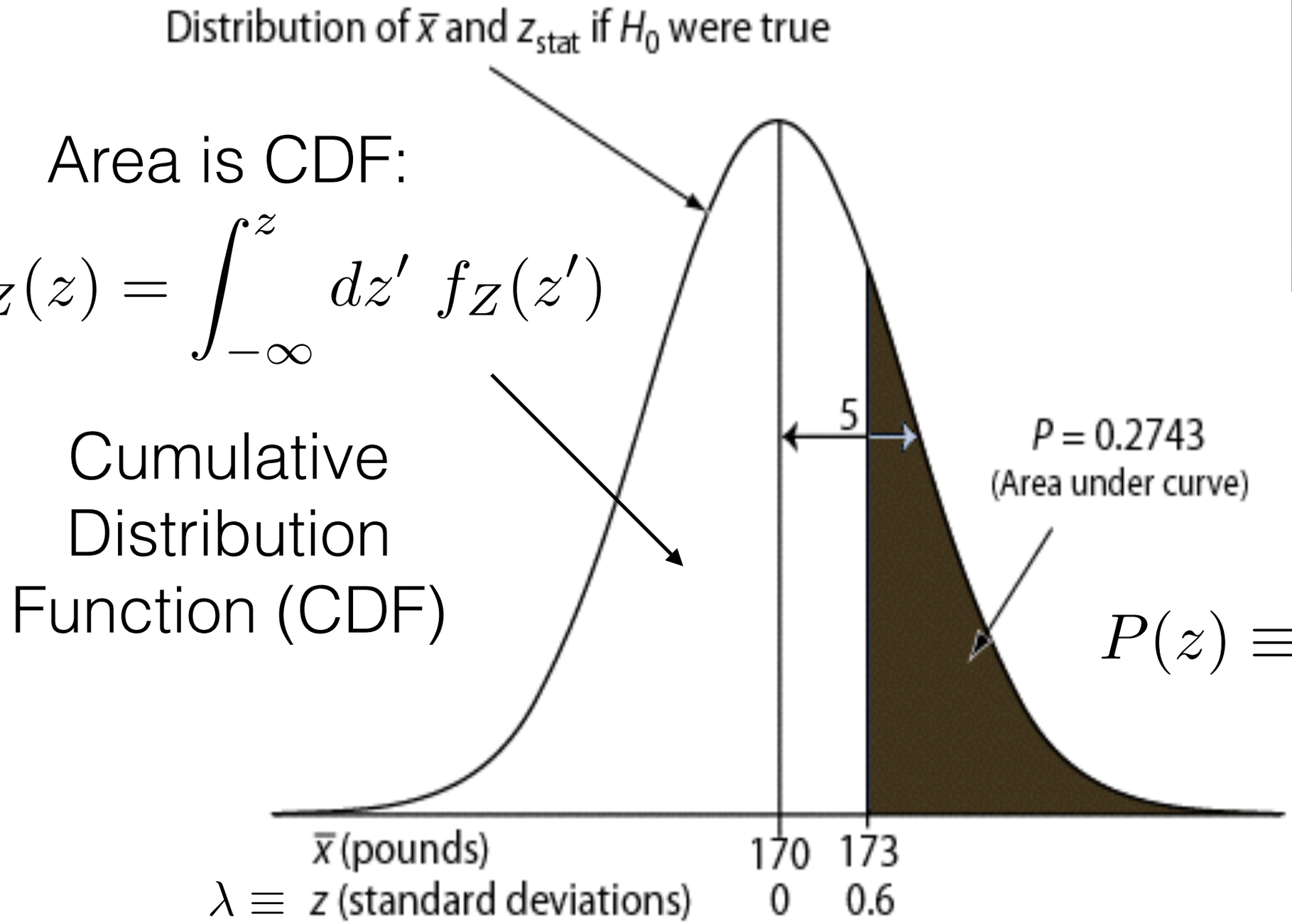
- The P-value is defined: $P(z) \equiv 1 - F_Z(z)$

P-value & the CDF of the test statistic

Area is CDF:

$$F_Z(z) = \int_{-\infty}^{z} dz' \; f_Z(z')$$

Cumulative Distribution Function (CDF)



R. A. Fisher

$$P(z) \equiv 1 - F_Z(z)$$

The *P*-value for Weight:

- How is the test statistic distributed?

$$\overline{X} \sim \mathcal{N}(\mu, \sigma_{\overline{x}}) \qquad \overline{Z} \equiv \frac{\overline{X} - \mu}{\sigma_{\overline{x}}} \sim \mathcal{N}(0, 1)$$

- Compute the Cumulative Probability Distribution (CDF) of the test statistic assuming the null hypothesis:

$$F_Z(z) = \tfrac{1}{2}[1 + \mathrm{erf}(z/\sqrt{2})]$$

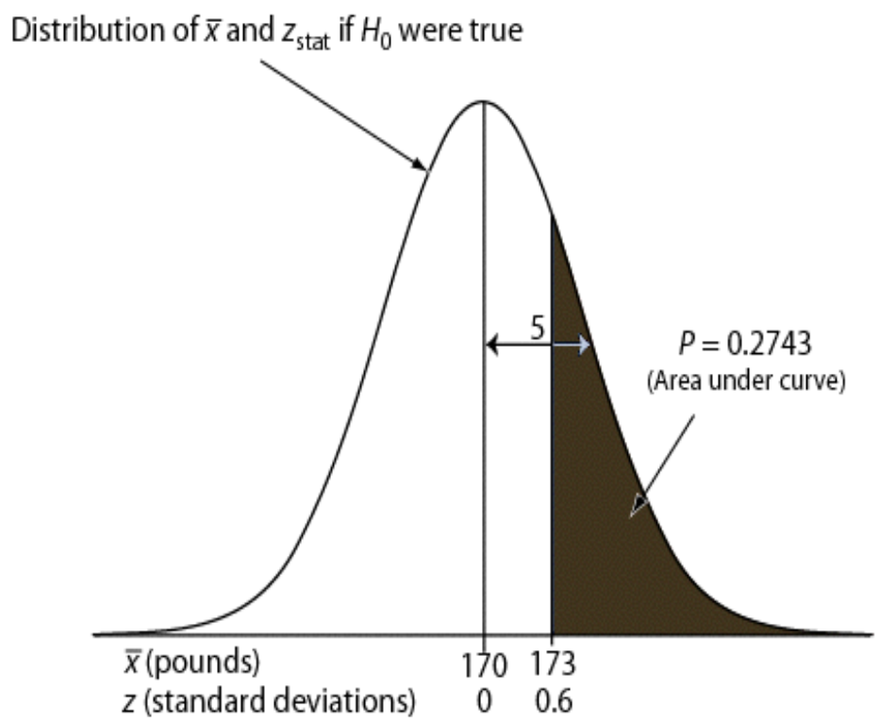
- The P-value is defined:

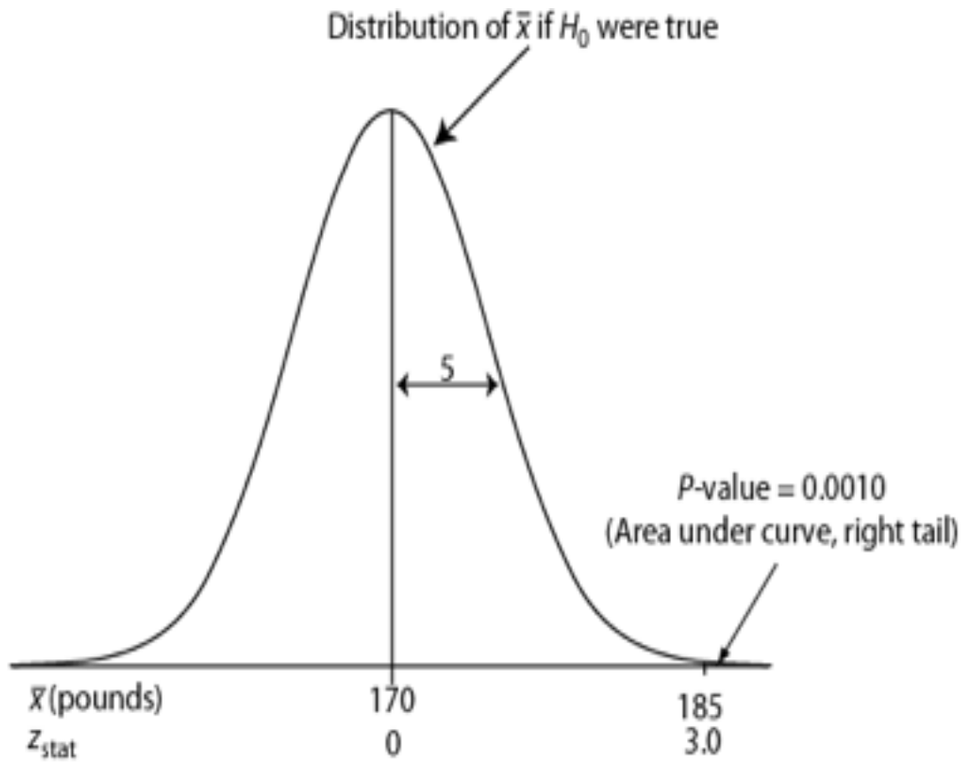
$$P(z) = 1 - F_Z(z)$$

- Where erf is the “error function”.

$$P(z) = \tfrac{1}{2}[1 - \mathrm{erf}(z/\sqrt{2})]$$

P-value for z_{stat} of 0.6 & 3.0





Interpretation

- *P*-value answer the question: What is the probability of the observed test statistic … **when H_0 *is true*?**
- Thus, smaller and smaller *P*-values provide stronger and stronger evidence against H_0
- Small *P*-value $\Rightarrow$ strong evidence
- P is **NOT** the probability of H_0

“Fisher” Interpretation

Conventions*

$P > 0.10 \Rightarrow$ non-significant evidence against H_0

$0.05 < P \leq 0.10 \Rightarrow$ marginally significant evidence

$0.01 < P \leq 0.05 \Rightarrow$ significant evidence against H_0

$P \leq 0.01 \Rightarrow$ highly significant evidence against H_0

Examples

$P = .27 \Rightarrow$ non-significant evidence against H_0

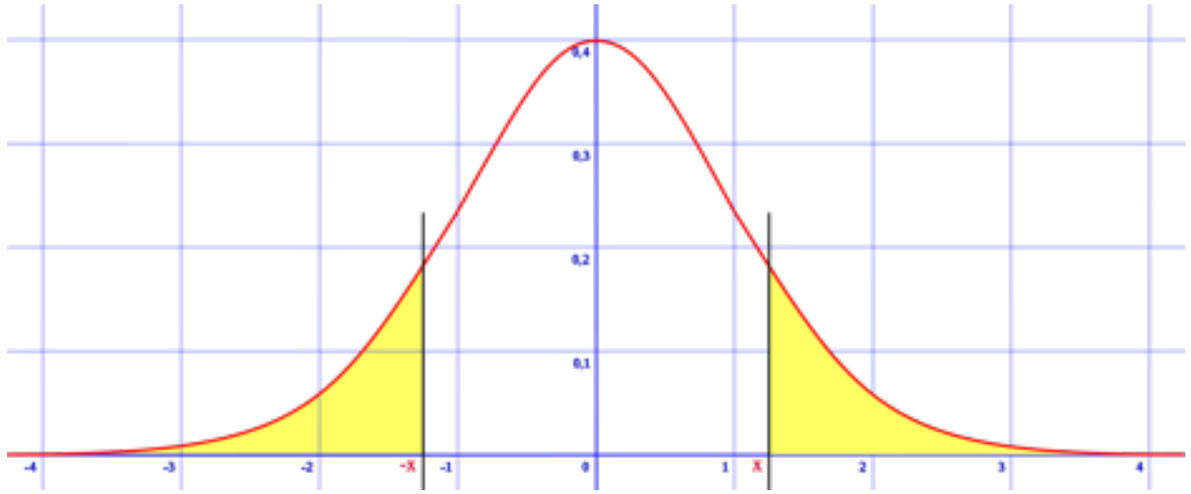
$P = .01 \Rightarrow$ highly significant evidence against H_0

*** It is *unwise* to draw firm borders for “significance”**

Where does the 2σ rule come from?

- My result is more the 2σ from the mean…

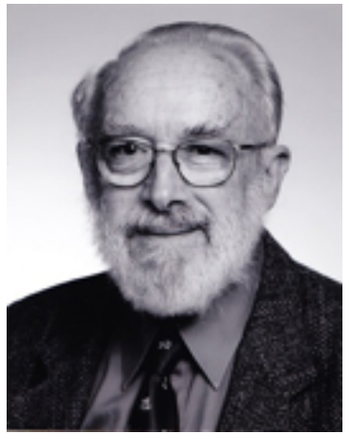
- Two-tailed p-value at 95% confidence level:



$$P(z) = 1 - \mathrm{erf}(z/\sqrt{2})$$

- Solve for: $P(z_*) = 1 - 0.95$

- Critical Z score: $z_* = 1.96 \approx 2$


D. Lindley

Is frequentism **really** *ad hoc*?

- Which test statistic is it that I need to use?
- Why does the P-value make sense (two-tailed, one-tail, …)?
- Why not look at PDF $f_Z(z)$ not the CDF $F_Z(z)$
- A different rule or method for every problem?

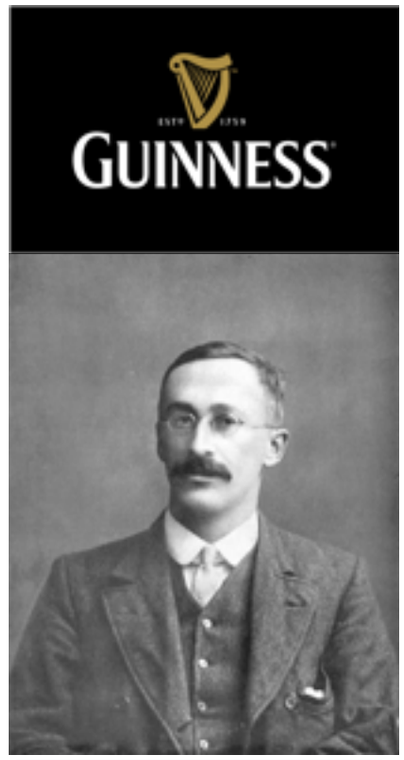
From the vault… some "**pre-packaged**" frequentist tests:

- Student's/Welch's t-test (pair/unpaired):
  "Are the unknown means equal is a normal model" (unknown equal/unequal variance).

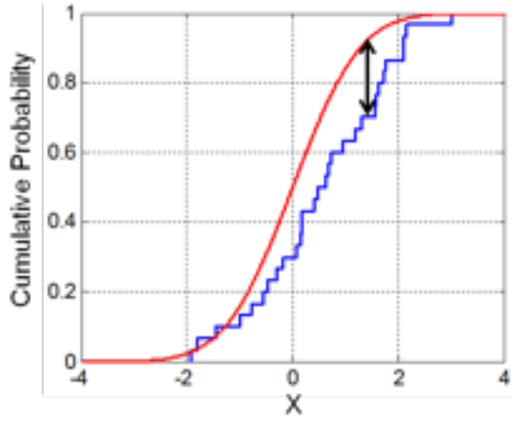
- but… what about the F-test, Z-test, *etc?*

- Kolmogorov–Smirnov test:
  Compare 1D PDFs…

- Weakness: *Difficult* to understand assumptions, weakness and which to use.



William Sealy Gosset



A **general** approach …
to choosing a **test statistic**

- Model = Candidate distributions + set of allowed parameter values:

$$q(x|\theta), \theta \in \Theta$$

- q is the model **likelihood**
- θ is the model **parameter**
- Examples:
  - Normal model with unknown mean:

$$q(x|\theta) \equiv (2\pi\sigma^2)^{-1/2} \exp[-\tfrac{1}{2\sigma^2}(x-\mu)^2], \theta \equiv (\mu)$$

  - Normal model with unknown mean and variance:

$$q(x|\theta) \equiv (2\pi\sigma^2)^{-1/2} \exp[-\tfrac{1}{2\sigma^2}(x-\mu)^2], \theta \equiv (\mu, \sigma^2)$$

Parameter estimation by **Maximum Likelihood Estimation**

- Meaning of Likelihood: Probability of the data given the parameters

- Estimate the parameters by maximizing the likelihood:

$$\hat{\theta} \equiv \arg\max_{\theta} q(x|\theta)$$

- Example of the normal model…

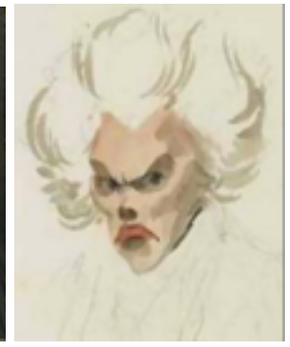
Show on board:

$$\hat{\mu} = \overline{X} \qquad \hat{\sigma^2} = \overline{(X - \overline{X})^2}$$

Population mean
μ = 170 pounds

BTW: You are already using MLEs…

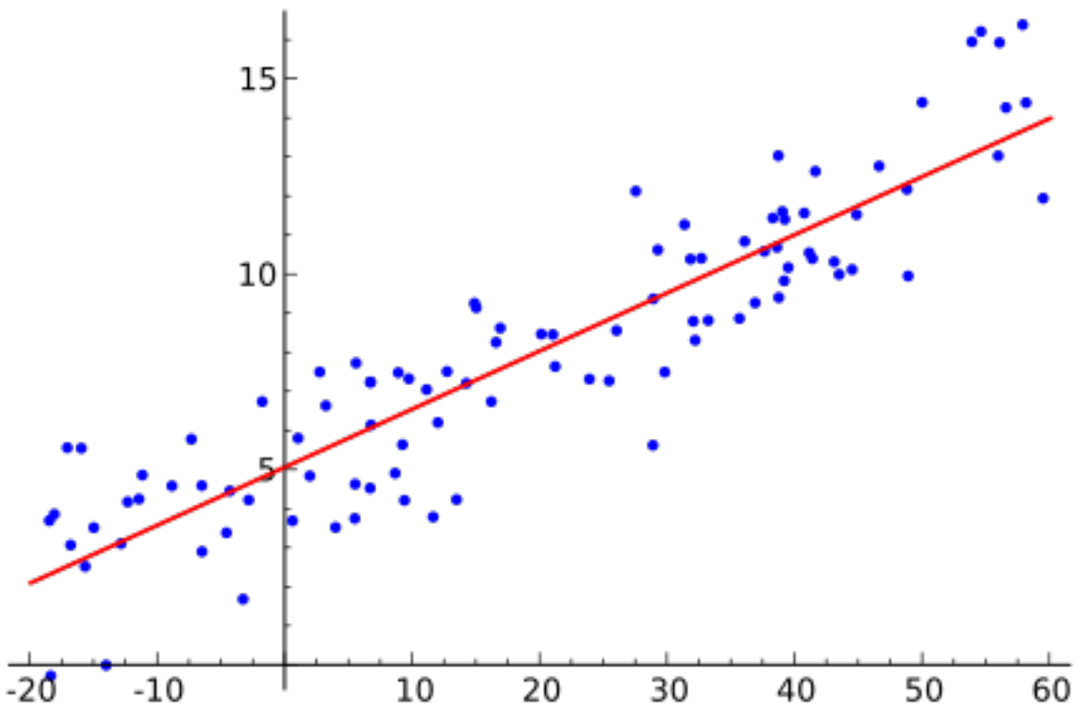
Carl Friedrich Gauß



Adrien-Marie Legendre

- Least Squares fitting IS the MLE procedure for a normal model with unknown mean

$$q(x|\theta) \equiv (2\pi\sigma^2)^{-1/2} \exp[-\tfrac{1}{2\sigma^2}(x-\mu)^2], \theta \equiv (\mu)$$



Claude Shannon

Shannon information:

$$h(x|\theta) = -\log q(x|\theta)$$

Show “min info” on board…

BTW: MLEs have some problems…

- At finite sample size, MLEs are typically biased.
    - Consider the ML estimator of variance:

Name:	Formula:	Expectation:
MLE estimator	$\hat{\sigma^2} = \overline{(X - \overline{X})^2}$	$\left\langle \hat{\sigma^2} \right\rangle = \frac{N-1}{N}\sigma^2$
Unbiased estimator	$\hat{\sigma^2} = \frac{N}{N-1}\overline{(X - \overline{X})^2}$	$\left\langle \hat{\sigma^2} \right\rangle = \sigma^2$

BTW: MLEs can be *fragile*: **German tank problem**

- Uniform distribution with unknown end-point:

$$X \sim \mathcal{U}(0, L)$$

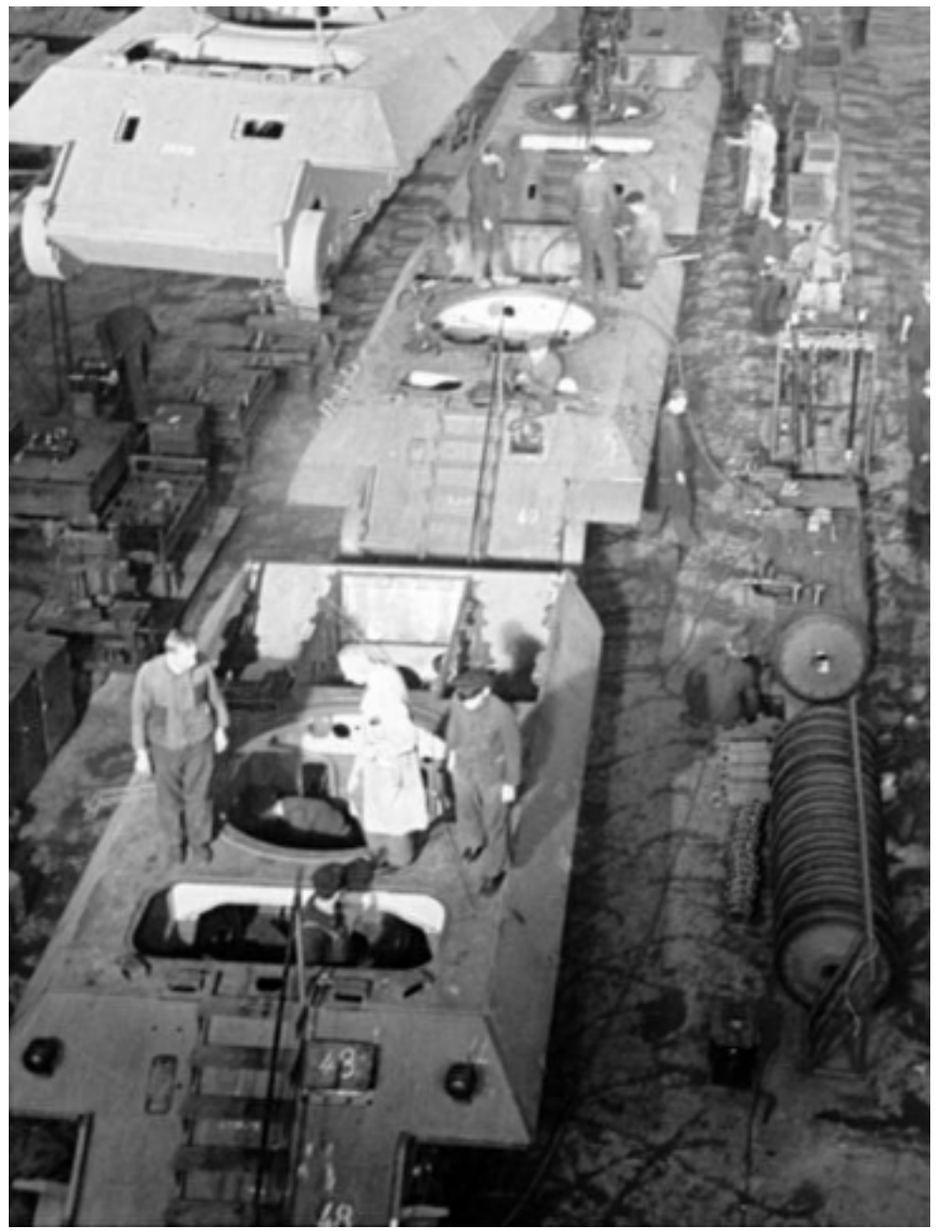
- Maximum likelihood Estimate:

$$\hat{L} = \max_{i=1\ldots N} \{X_i\}$$

- Stupid estimate (why?)

- Use Bayesian inference…

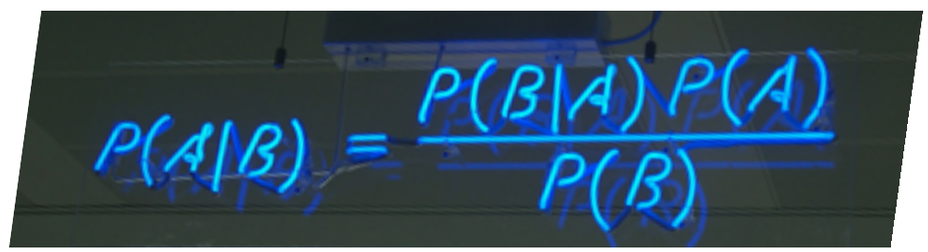
Month	Statistical estimate	Intelligence estimate	German records
June 1940	169	1,000	122
June 1941	244	1,550	271
August 1942	327	1,550	342



P. S. Laplace

T. Bayes

A different approach: Bayesian inference…



- Joint probability: $p(x, \theta)$

- Marginalized Probabilities:

$$p(\theta) = \sum_x p(x) \qquad p(x) = \sum_\theta p(x, \theta)$$

- Conditional probabilities:

$$p(\theta|x) \equiv p(x, \theta)/p(x) \qquad p(x|\theta) \equiv p(x, \theta)/p(\theta)$$

- Bayes' Theorem:

$$p(\theta|x) = p(x|\theta)p(\theta)/p(x) \qquad p(x|\theta) = p(\theta|x)p(x)/p(\theta)$$

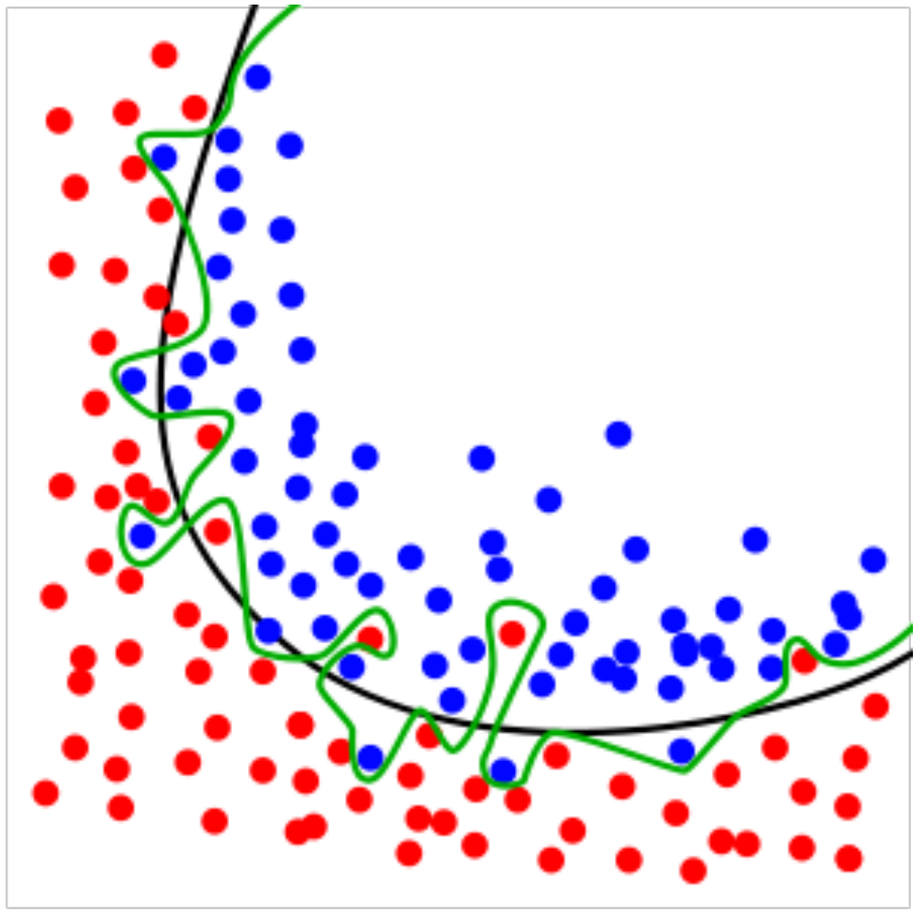
The general approach to **Null hypothesis testing**

- Meaning of Likelihood: Probability of the data given the parameters

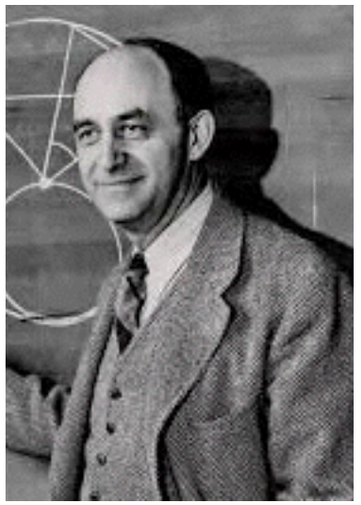
- Find the model that maximizes the likelihood

- “Biassed estimator”: more parameters **ALWAYS** lead to a better fit

- Solution: **Likelihood Ratio Test**

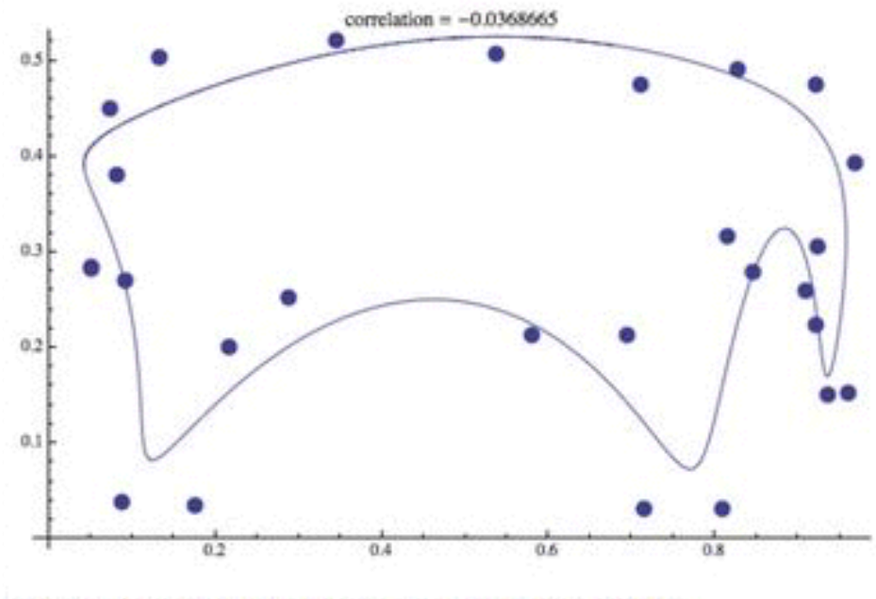


The curse of dimensionality (1)



E. Fermi

F. Dyson



- Fermi to Dyson: “With **4 parameters** I can fit an **elephant**, and with **5** I can make him **wiggle his trunk.**”

- **Hughes effect:** The predictive **performance decreases** as model **dimension increases** due to **overfitting**

Richard E. Bellman & Gordon F. Hughes

The general approach to **Null hypothesis testing**

- Log Likelihood Ratio test statistic:

$$\lambda(X^N) \equiv \log q_1(X^N|\hat{\theta})/q_0(X^N|\hat{\theta})$$

- Assume null hypothesis and compute the distribution of the test statistic:

$$\Lambda \sim \lambda(X^N) \quad \text{where} \qquad X \sim H_0$$

- Choose a significance level α: (95% confidence is α = 1-0.95 = 0.05)

- Compute the P-value (F is Cumulative Distribution Function) :

$$p(\lambda) = 1 - F_\Lambda(\lambda)$$

- Accept ($\alpha < p(\lambda)$) or reject null hypothesis ($\alpha > p(\lambda)$)

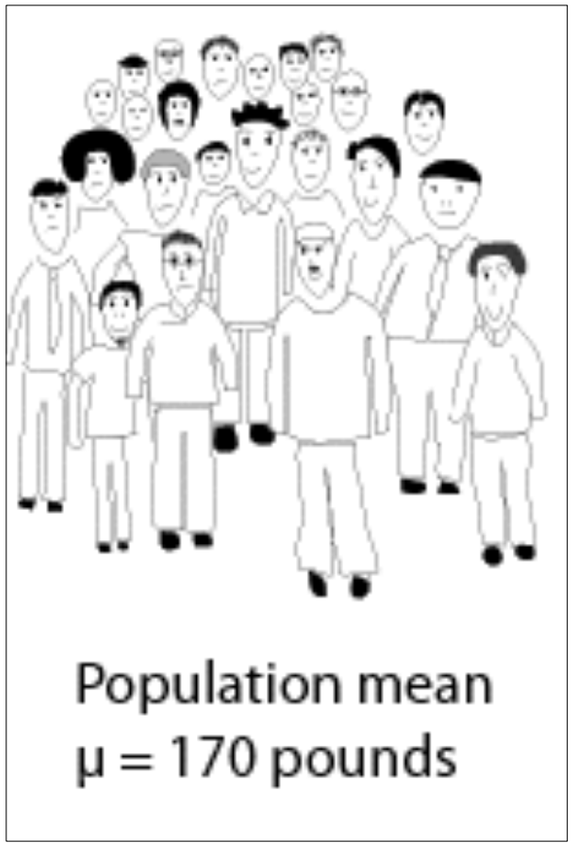
Example: Likelihood Ratio Test

- Null hypothesis:

$$\mu = 0, \sigma^2 = 1$$

- Alternative hypothesis:

$$\mu \neq 0, \sigma^2 = 1$$



- Show that the test statistics is distributed like a 1D chi-squared distribution:

$$\Lambda \sim \frac{1}{2}\chi_1^2$$

Wilks' theorem:

Samuel S. Wilks

- “Wow… that is a lot of work even for a normal model. How will I ever compute this for my model?”

- Monte Carlo simulation

- Wilks’ theorem: Large-sample-size limit of the likelihood ratio test (for a nested, regular model):

$$\lim_{N \to \infty} \Lambda \sim \tfrac{1}{2} \chi^2_k$$

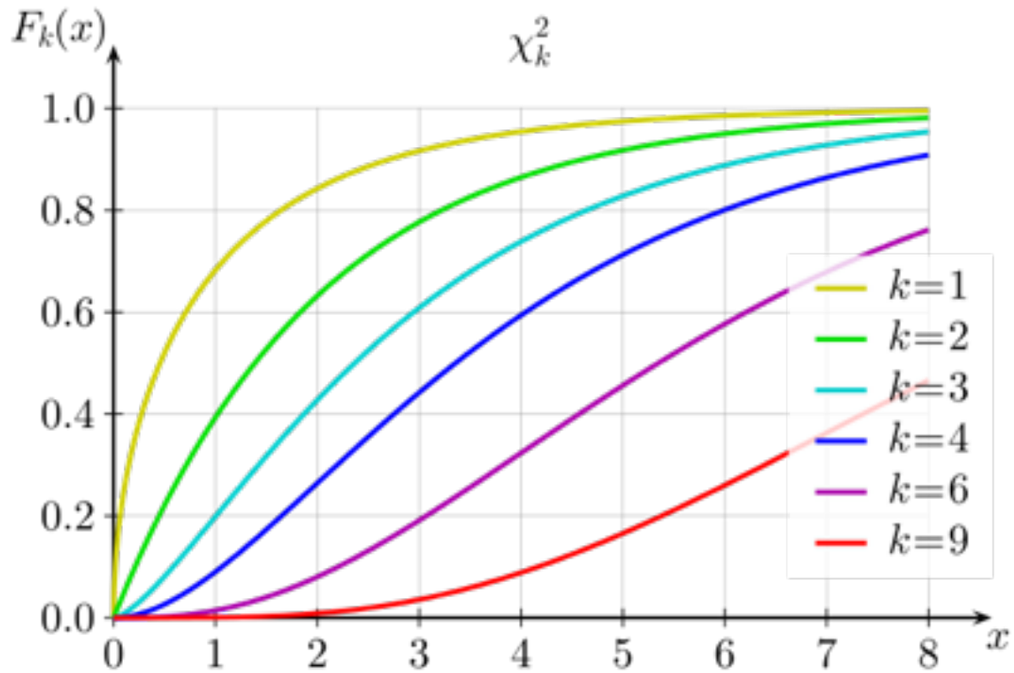
k = difference in dimension between alternative and null model

Example: Likelihood Ratio Test

- Compute the P-value:
  - Compute CDF:

$$F_k(\lambda) \equiv \Gamma(k/2)^{-1}\gamma(k/2, \lambda)$$

Incomplete gamma function (k = 1).



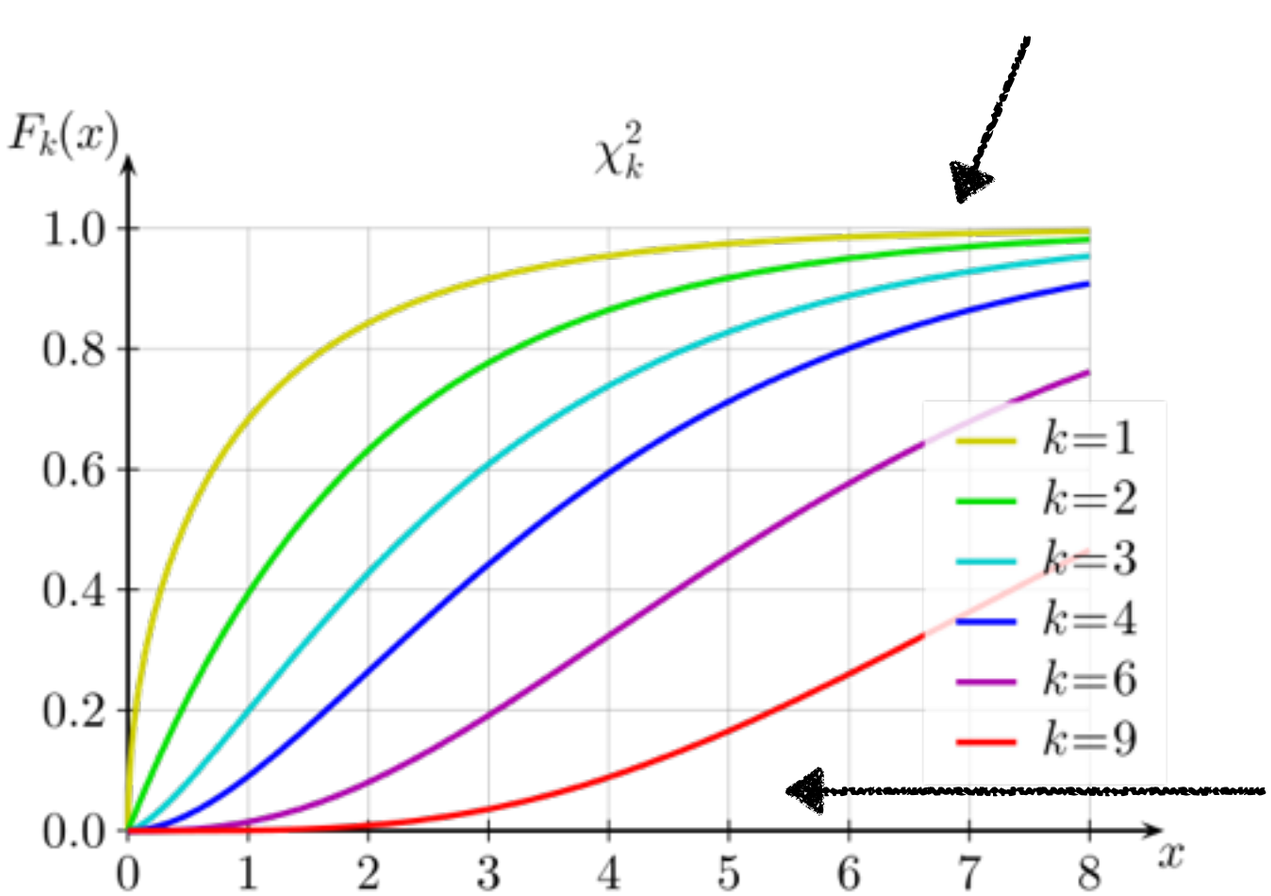
- Compute P-value:

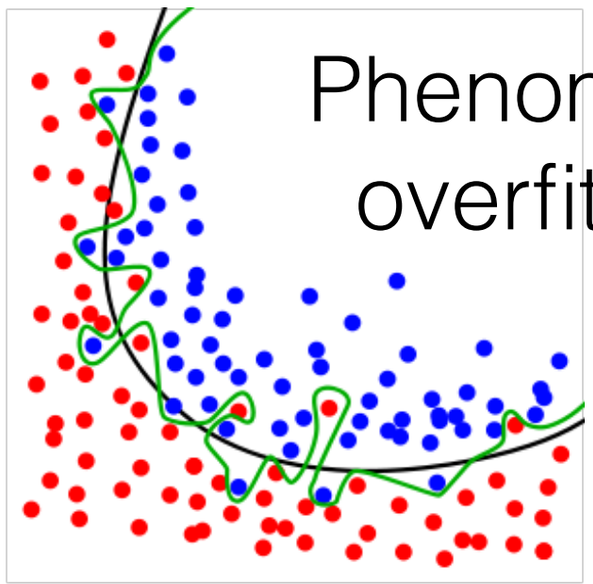
$$P(\lambda) = 1 - F_k(\lambda)$$

The curse of dimensionality (2)

William of Ockham

H_1 dimension is **small**… “less evidence” required





Phenomena of overfitting…

H_1 dimension is **big**… “more evidence” required