3011979 Practical Python for Data Sciences and Machine Learning

L4: Statistics for data science and ML

Feb 4th, 2022

Sira Sriswasdi, Ph.D.
Research Affairs, Faculty of Medicine
Chulalongkorn University

Probability

Why probability?

- Statistics relies on probability
- P-value = **probability** of observing the same or more extreme result, given that the null hypothesis is true
- Model that best fits the data is the one that maximize P(data | **model**) ~ maximum likelihood estimator
- Model that best fits the data is the one that maximize P(**model** | data) ~ maximum a posteriori probability

Probability lets you model the world

- In Bangkok with 7 million population and a daily new COVID-19 case of 7,000, what's the probability that you will be infected tomorrow?

- Consider a gene with 2 alleles, *NFT* and *nft*. If the frequency of *NFT* in Thailand is 0.8, what is the probability that you have genotype *NFT* / *NFT*. What about *NFT* / *nft* ?

- If the chance of death from ASF is 60% each day, what is the probability that an infected pig will survive 5 days?

Key terminology

- **Conditional probability**: P(A | B)
  - Probability that you have genotype *NFT* / *NFT* **given** that the frequency of *NFT* is 0.8
  - Probability that you have genotype *NFT* / *NFT* **given** that the frequency of *NFT* is 0.8 and that the genotype *nft* / *nft* is lethal

- **Joint probability**: P(A, B) = P(A | B) P (B)
  - Probability of observing A and B = probability of observing B **first**, followed by observing A **given** that B has already occurred
  - Probability that you have genotype *NFT* / *NFT* **and** that the frequency of *NFT* is 0.8

- **Bayes' rule**: P(A | B) = P(B | A) x P(A) / P(B)

Some useful probability distribution

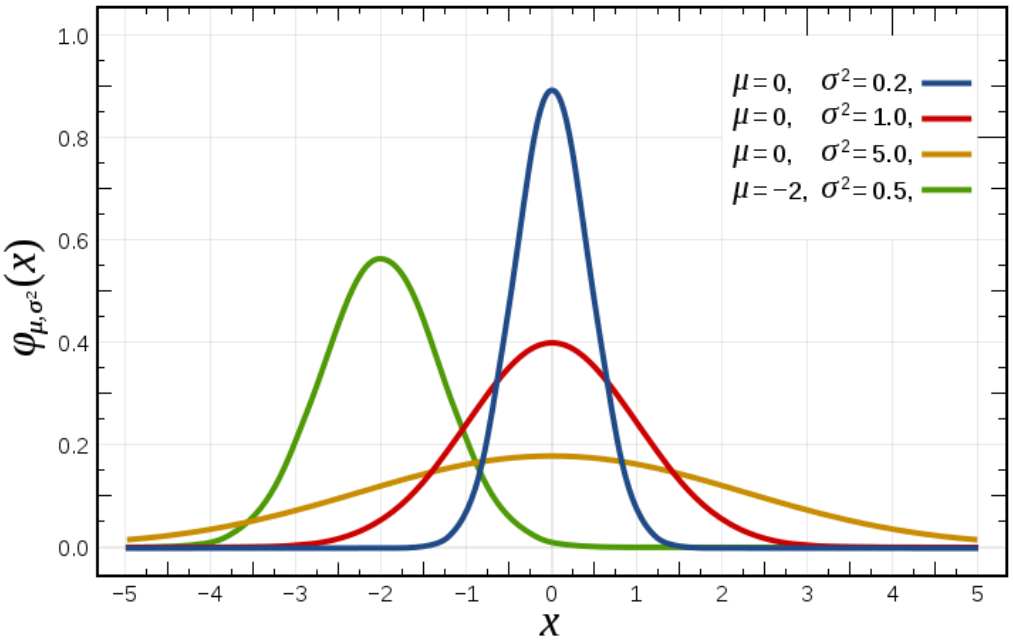
- Binomial
  - A process with two outcomes: *Win* and *Lose*
  - P(*Win*) = *p*, P(*Lose*) = 1 – *p*
  - What is the probability of getting *k* wins out of *n* trials?
  - What is the expected value?

- If model A and B fits the data equally well, what is the probability that model A achieves higher accuracy that model B in 80 out of 100 evaluations?

- If feature X is completely unrelated to treatment outcome, what is the probability that the coefficients of X in 100 logistic regression models all be positive?

Normal distribution

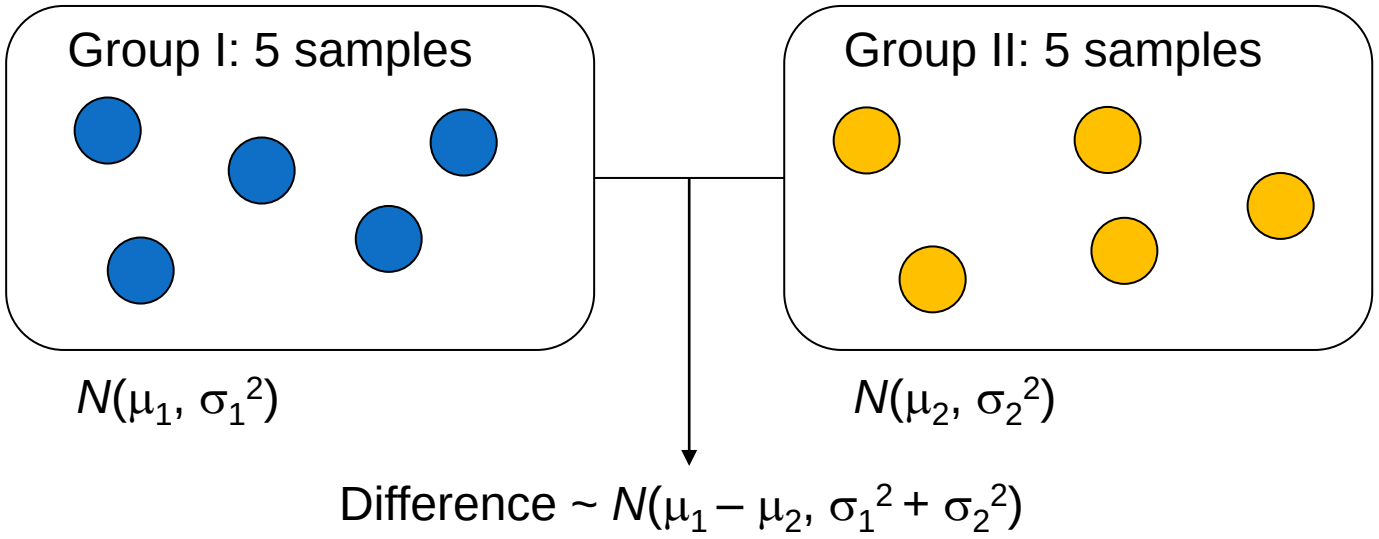
- Also called Gaussian distribution
- Bell-shaped and defined by two parameters
  - Location: Mean μ
  - Scale: Variance σ^2



Images from https://en.wikipedia.org/wiki/Normal_distribution

Properties of normal distribution

- Closed under addition and scalar multiplication
  - $N(\mu, \sigma^2) + a \rightarrow N(\mu + a, \sigma^2)$
  - $N(\mu, \sigma^2) * a \rightarrow N(\mu * a, \sigma^2 * a^2)$
  - $N(\mu_1, \sigma_1^2) + N(\mu_2, \sigma_2^2) \rightarrow N(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$

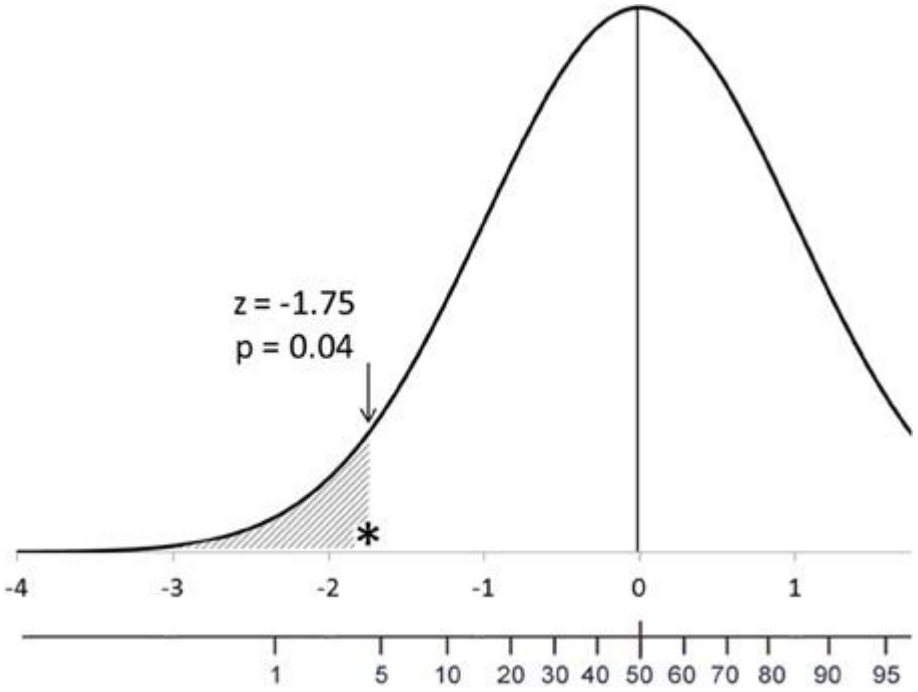


- Easy to assess the likelihood of observing difference larger than some value or smaller than some value

Standard normal distribution

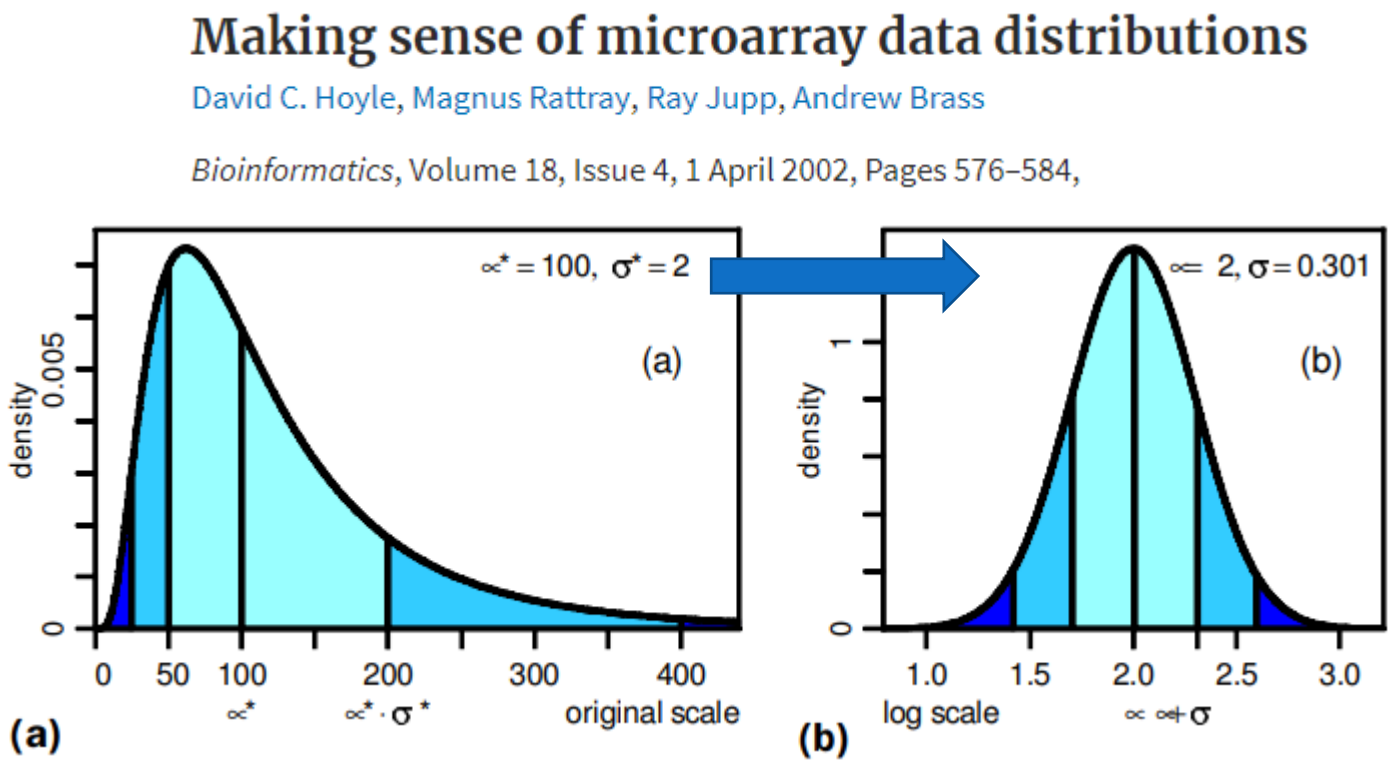
- Normal distribution with zero mean and unit variance
- How to transform data from a normal distribution with mean μ and variance σ^2 to a standard normal distribution?

- Z-score



Log-normal distribution

- Distribution whose log-values are normally distributed
- Value ranges from 0 to ∞
- Intensity data, such as microarray's and mass spectrometry's



Limpert, Stahel, and Abbt. BioScience 2001.

Chi-square distribution

If $Z_1, \ldots, Z_k$ are independent, standard normal random variables, then the sum of their squares,

$$Q = \sum_{i=1}^{k} Z_i^2,$$

is distributed according to the chi-square distribution with k degrees of freedom. This is usually denoted as

$$Q \sim \chi^2(k) \text{ or } Q \sim \chi_k^2.$$

The chi-square distribution has one parameter: a positive integer k that specifies the number of degrees of freedom (the number of Z_i s).

- Test for categorical observation $\chi^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i}$
  - Whether the observed distribution followed expectation
  - Association between two categorical features

Statistics

Statistics

- Quantify uncertainty
  - How much can you trust the observed data?
  - Is the conclusion robust to small perturbation?

- Assess models and data
  - Which model family better fit the data?
  - What are the model parameters that best explain the data?
  - Are data from multiple experiments comparable?
  - Is the data normally distributed?

Describing data average

- Arithmetic mean: (x + y) / 2

- Geometric mean: sqrt(x * y) = exp(log(x) + log(y) / 2)
  - Log of GM = AM of Log
  - Good for exponential data
  - Lean toward the smaller value

- Harmonic mean: 2 / (1/x + 1/y) = 1 / { (1/x + 1/y) / 2 }
  - Inverse of HM = AM of Inverse
  - Lean even more toward the smaller value

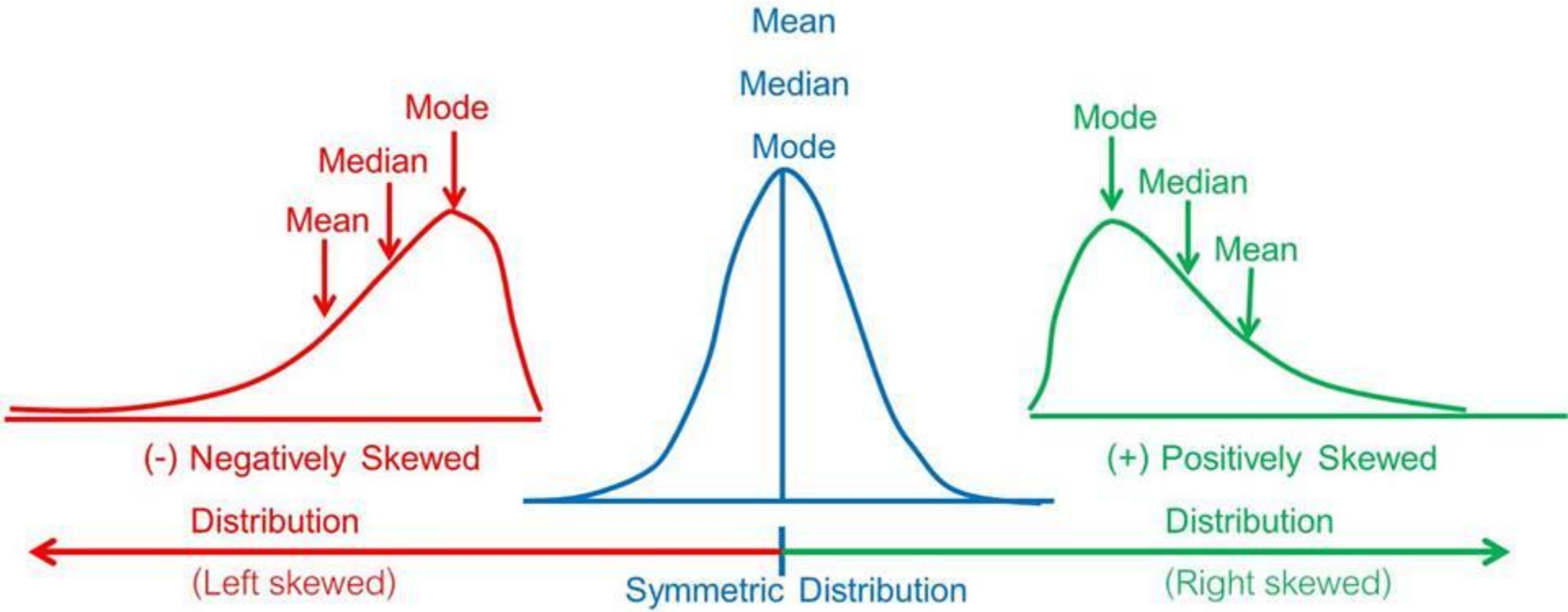
- Median & Mode

Moments

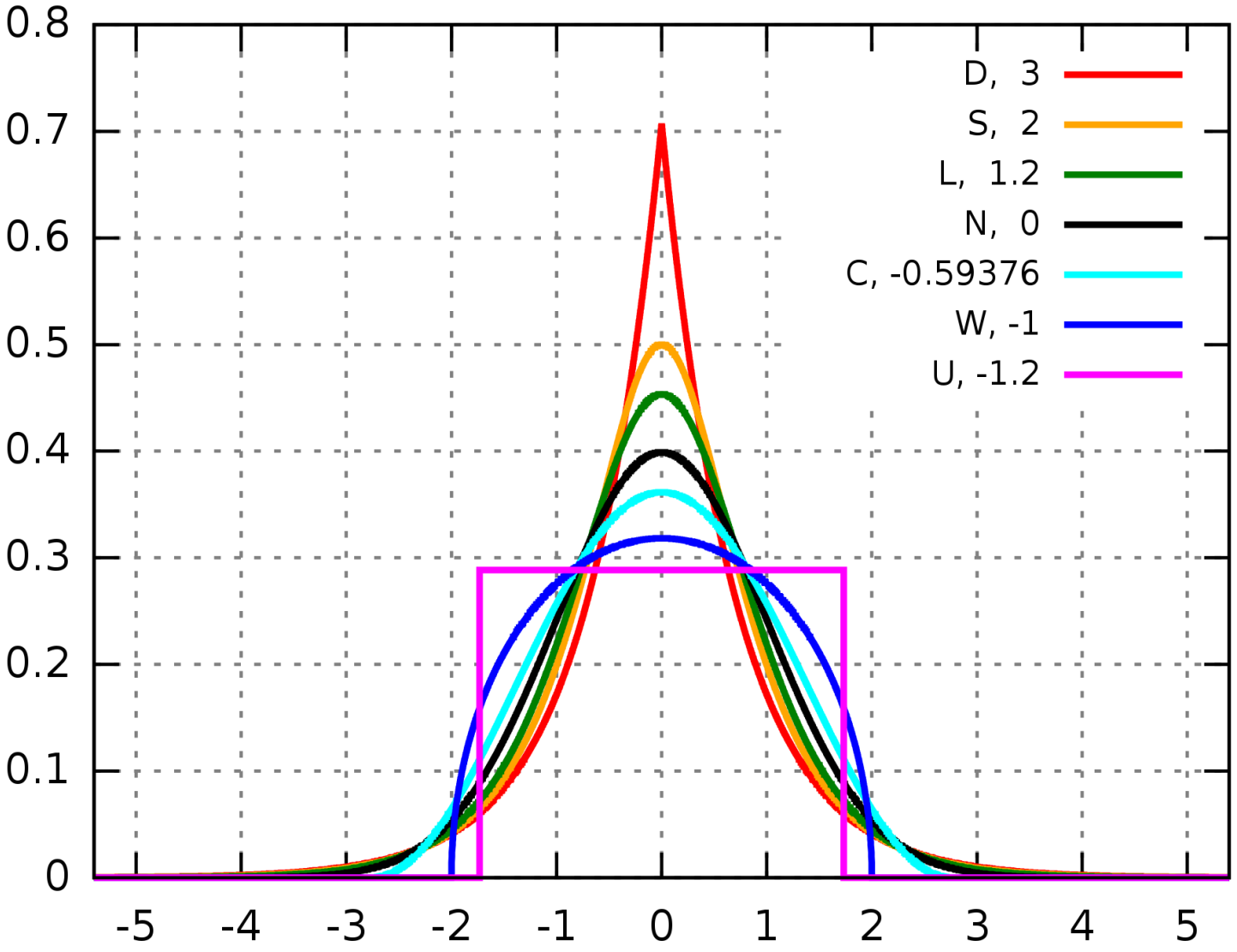
Moment ordinal	Moment		
	Raw	Central	Standardized
1	Mean	0	0
2	–	Variance	1
3	–	–	Skewness
4	–	–	(Non-excess or historical) kurtosis

- Raw moment = $E[X^n]$
  - Mean = E[X], expected value of X
- Central moment = $E[(X - \mu)^n]$
  - Variance = $E[(X - \mu)^2] = E[X^2] - E[X]^2$
- Standardized moment = $E[(X - \mu)^n] / \sigma^n$
  - Skewness ~ symmetry (n = 3)
  - Kurtosis ~ heaviness of the tail (n = 4)

Skewness


Mean
Median
Mode
Mode
Median
Mean
Mode
Median
Mean
(-) Negatively Skewed
Distribution
(Left skewed)
Symmetric Distribution
(+) Positively Skewed
Distribution
(Right skewed)

Kurtosis



Estimate the model parameters

- In Bangkok with 7 million population and a daily new COVID-19 case of 7,000, what is the infection rate?

- If 18 out of 1,000 person have genotype *NFT* / *NFT*. What is the frequency of *NFT* allele in the population?

- If 67 out of 1,000 pigs with ASF died on the first day and 541 more died on the second day, what is the rate of death each day for ASF?

Inferring model parameters

- X = observed data
- $\boldsymbol{\theta}$ = model parameters
- Maximum Likelihood Estimator (MLE)
  - Likelihood $= \mathrm{P}(\mathrm{data} \mid \mathrm{model}) = \mathrm{P}(X \mid \boldsymbol{\theta})$
  - $\boldsymbol{\theta}_{\mathrm{MLE}} = \mathrm{argmax}\ \mathrm{P}(X \mid \boldsymbol{\theta})$
    - Find $\boldsymbol{\theta}$ that maximize the probability of observing the data
    - P(a patient survive for 5 year | yearly survival rate = $\boldsymbol{\theta}$)
    - P(a cohort of patients survive for [5, 4, 10, 7, 3, 1] year | yearly survival rate = $\boldsymbol{\theta}$)

Hypothesis testing

Deciding between possibilities

- Is coefficient of a feature positive or negative?
- Is survival rate equal to 0.1 or 0.2 or …?
- Is infection rate greater than 0.5?

- These possibilities are **hypotheses**

Null and alternative hypotheses

- Often, the likelihood of the model you are interested in cannot be calculated (**alternative hypothesis**):
  - Average yearly survival rate of patient = 0.5
  - We want to assess the benefit of a new treatment and found that a cohort of patients seem to survive longer
  - P(the new treatment is more effective than traditional ones)
  - P(a cohort of patients survive for [5, 4, 10, 7, 3, 1] year | yearly survival rate is greater than 0.5)

- Instead, the likelihood of the default model is much easier to calculate (**null hypothesis**):
  - P(a cohort of patients survive for [5, 4, 10, 7, 3, 1] year | yearly survival rate is exactly 0.5)

If likelihoods are computable:

- Likelihood ratio test
  - Log Likelihood Ratio (LLR) = Log[P(X | $\boldsymbol{\theta}_1$) / P(X | $\boldsymbol{\theta}_2$)]
  - Reject $\boldsymbol{\theta}_2$ if LLR > k
  - Reject $\boldsymbol{\theta}_1$ if LLR < $-k$
  - This is a **most powerful** test (Neyman-Pearson Lemma)

- Nested model test
  - Compare models within the same family, one is nested inside another
    - Model 1: Patient survival depends on age
    - Model 2: Patient survival depends on age and drug choice
  - Test for significant difference in likelihood P(X | $\boldsymbol{\theta}_2$) – P(X | $\boldsymbol{\theta}_1$)
  - Related to Chi-squared distribution

- Akaike Information Criterion
  - AIC = 2 x number of parameters – log likelihood

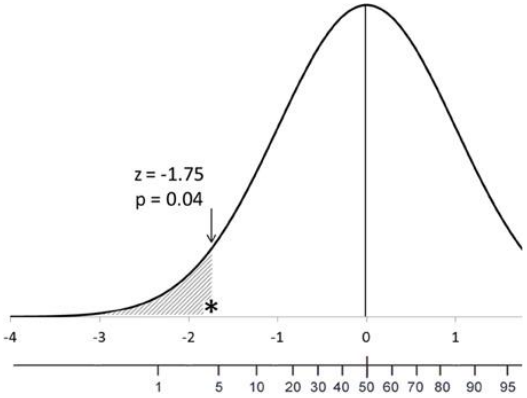
Otherwise, we rely on only p-value

- Start with an alternative hypothesis
  - Inspect the data to spot interesting patterns using domain knowledge
  - Our patient cohort survives longer than normal

- Identify the appropriate null hypothesis
  - What would be expected by chance?
  - Survival rate of past patients

- P-value = probability of observing the same or more extreme result, given that the **null hypothesis** is true
  - Rely on only the null hypothesis
  - Reject null hypothesis if p-value is small

Hypothesis testing framework



- Propose null and alternative hypotheses
  - The sample mean is equal to μ_0
- Decide on the test to use and compute **test statistic *t***
  - One-sample test with test statistics $t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$
- Derive the distribution of ***t*** under the null hypothesis
  - Sample mean is normally distributed with large number of sample
  - ***t*** ~ N(0, 1) by Central Limit Theorem
- Specify the **significance level** α to reject null hypothesis
- Compute **p-value**: p($t > t_{obs}$ | null hypothesis)
  - Correspond to cdf of normal distribution

It's all about test statistics *t*

- One sample *t*-statistics: $t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$
- U statistics: $U = \sum_{i=1}^{n} \sum_{j=1}^{m} S(X_i, Y_j), \quad S(X,Y) = \begin{cases} 1, & \text{if } X > Y, \\ \frac{1}{2}, & \text{if } Y = X, \\ 0, & \text{if } X < Y. \end{cases}$
- Signed-rank statistics:

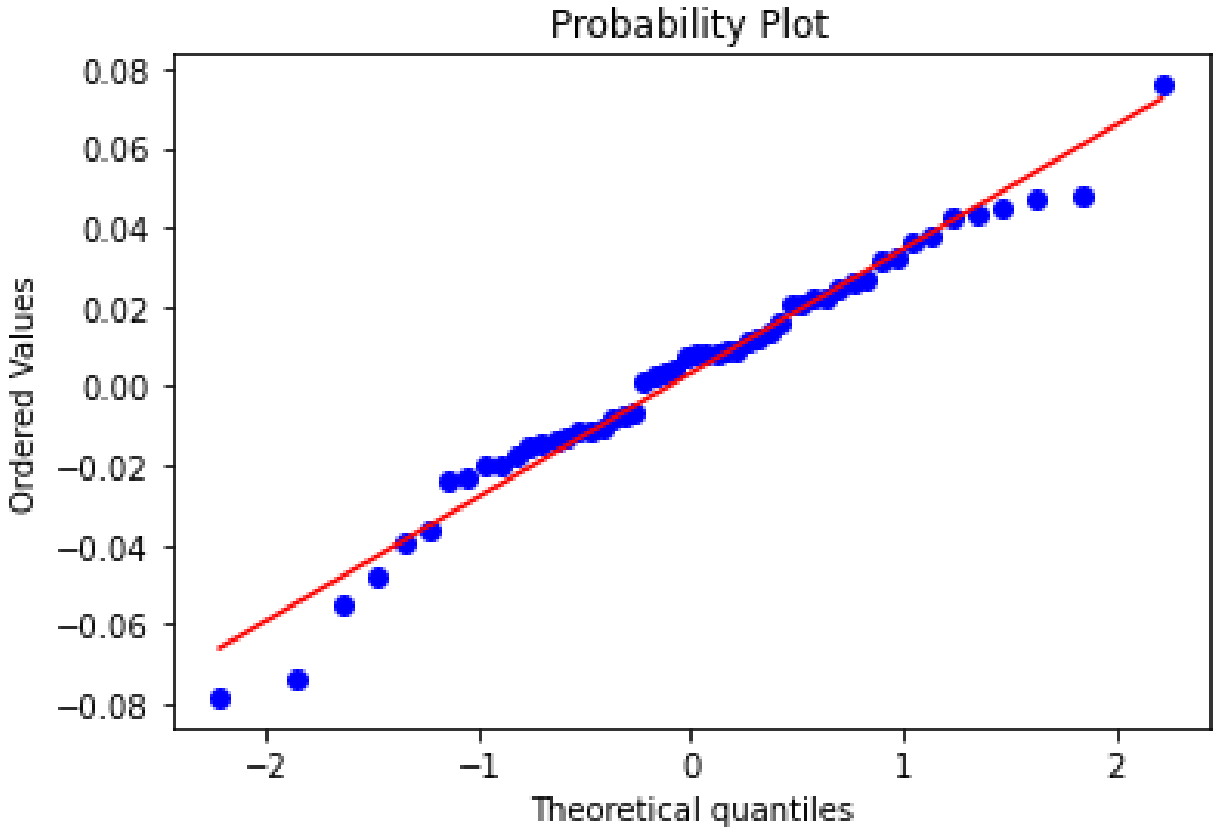
1. Compute $|X_1|, \ldots, |X_n|$.
2. Sort $|X_1|, \ldots, |X_n|$, and use this sorted list to assign ranks $R_1, \ldots, R_n$

$$T = \sum_{i=1}^{N} \operatorname{sgn}(X_i) R_i.$$

- Sign test: P(X > Y) = 0.5, with Binomial distribution
- And whether one can derive the distribution of these statistics under the null hypothesis...

Useful tests

Visual distribution test with Q-Q plot



- Compare observed versus theoretical quartiles
  - Estimated based on sample mean and variance
- Can be used for any distribution

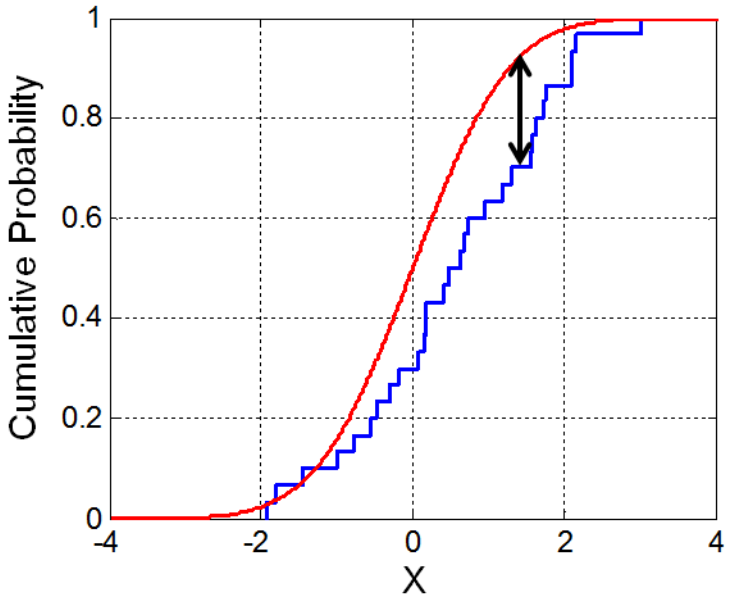
Tests of normality

- Jarque-Bera (*scipy.stats.jarque_bera*): $JB = \frac{n}{6} \times \left(S^2 + \frac{(K-3)^2}{4} \right)$

- Shapiro-Wilk (*scipy.stats.shapiro*): $W = \frac{(\sum_{i=1}^{n} a_i x_{(i)})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$

  where $x_{(i)}$ is the ith smallest number in the sample $(x_1 < x_2 < \ldots < x_n)$; and a_i are constants generated from var, cov, mean for a normally distributed sample.

- Kolmogorov-Smirnov (*scipy.stats.kstest*):



Kolmogorov distribution [edit]

The Kolmogorov distribution is the distribution of the random variable

$$K = \sup_{t \in [0,1]} |B(t)|$$

where $B(t)$ is the Brownian bridge. The cumulative distribution function of K is given by[2]

$$\Pr(K \leq x) = 1 - 2\sum_{k=1}^{\infty}(-1)^{k-1}e^{-2k^2x^2} = \frac{\sqrt{2\pi}}{x}\sum_{k=1}^{\infty}e^{-(2k-1)^2\pi^2/(8x^2)},$$

More tests of normality in *scipy*

(Too) Many options

- D'Agostino's K-squared test,
- Jarque–Bera test,
- Anderson–Darling test,
- Cramér–von Mises criterion,
- Kolmogorov–Smirnov test
- Lilliefors test based on the Kolmogorov–Smirnov test
- Shapiro–Wilk test
- Pearson's chi-squared test

- scipy.stats.normaltest
  - Based on D'Agostino's and Pearson's methods
  - Null hypothesis: Data is normally distributed
  - Test statistic: normalized skewness^2 + kurtosis^2
  - Distribution of test statistic: Chi-square

Tests for comparing samples

- Student's *t*-test (assuming normally distributed data)
  - *scipy.stats.ttest_1samp*: one data against a mean
  - *scipy.stats.ttest_ind*: two independent data
  - *scipy.stats.ttest_rel*: paired data

- Non-parametric test (rank-based)
  - *scipy.stats.mannwhitneyu*: two independent data
  - *scipy.stats.wilcoxon*: paired data

- Sign test (sign-based)
  - *statsmodels.stats.descriptivestats.sign_test*
  - Or you can calculate p-value using Binomial distribution
    - *scipy.stats.binom*

Impact of test choices

Data	Exp 1	Exp 2	Exp 3	Exp 4	Exp 5	Exp 6	Exp 7
Model A	0.701	0.503	0.991	0.827	0.623	0.728	0.596
Model B	0.691	0.478	0.905	0.739	0.589	0.719	0.508

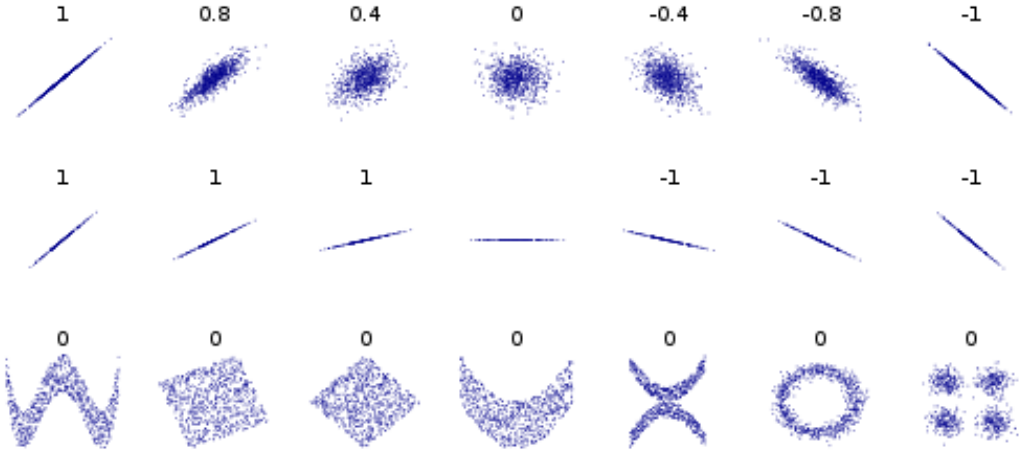
- Does model A perform better than B?
  - Unpaired Student's t-test p-value = 0.5687
  - Unpaired Mann-Whitney U test p-value = 0.6101
  - Paired Student's t-test p-value = 0.0137
  - Wilcoxon signed rank test p-value = 0.0156
  - Sign test p-value = 0.00815

Empirical hypothesis testing (Permutation test)

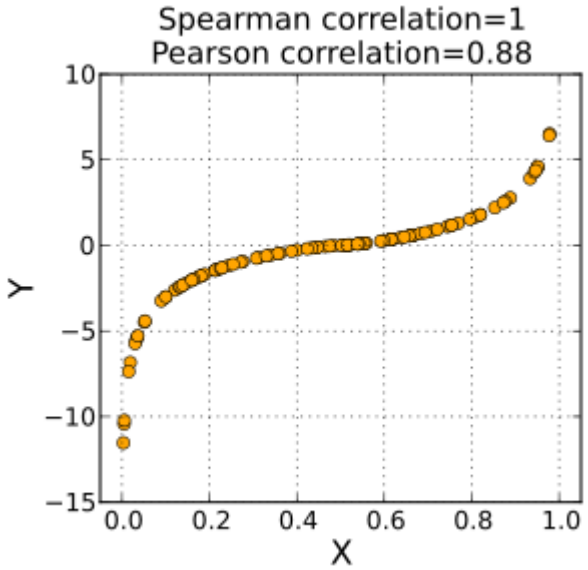
Correlation

- *scipy.stats.pearsonr* measures linear relationship
- *scipy.stats.spearmanr* measures rank relationship
- *scipy.stats.kendalltau* also measures rank

Pearson's correlation reflects ONLY linearity



Images from https://en.wikipedia.org/wiki/Correlation_and_dependence



Images from https://en.wikipedia.org/wiki/Spearman%27s_rank_correlation_coefficient

Significance of correlation

- Without assumption on the distribution of the data, it is **impossible** to calculate the distribution of correlation
  - *scipy*'s implementation assumes normal distribution

- Empirical p-value calculation with permutation test
  - **Alternative hypothesis**: High correlation = *r* due to association structure between the two data
  - **Null hypothesis**: There is no association structure
  - **Test statistics**: correlation coefficient
  - **Distribution of test statistics under null hypothesis**: Shuffle one data multiple times and recalculate correlation coefficient
  - **P-value**: Fraction of shuffles with correlation coefficient ≥ *r*

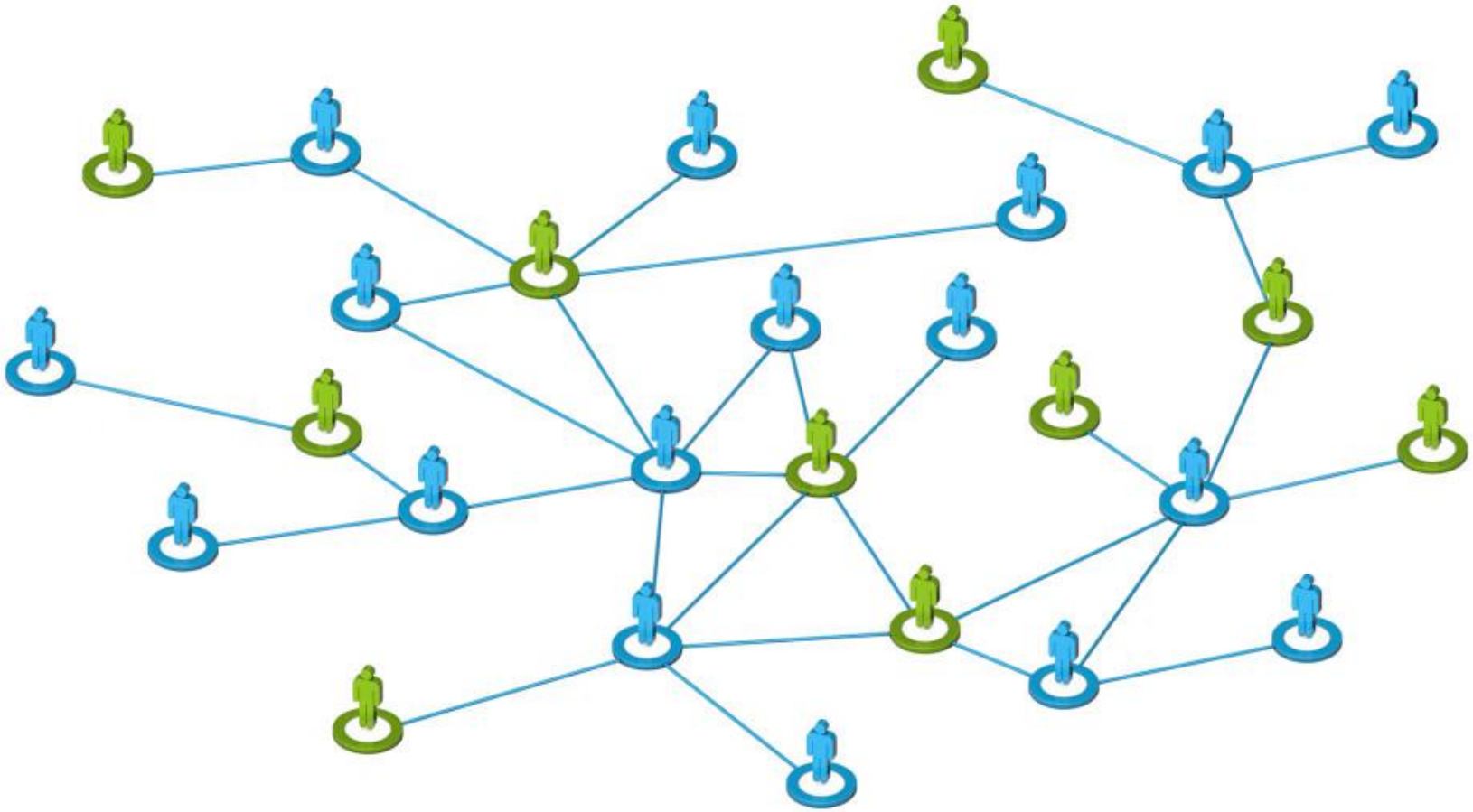
Permutation test

- Test the hypothesis that data have some structures
  - Significance of observed correlation
  - Model A outperforms model B

- Shuffle data multiple times and recompute the occurrence of patterns or statistics of interest
  - Count how many times more extreme values are obtained
  - Fraction of times = P-value

- Highly depend on the shuffling
  - Some structure in the data must be preserved
    - Height and weight
  - Otherwise, null hypothesis may be rejected due to the unintended missing structure

Network data



- Node = FB page, Edge = friendship
- Node color = gender
- **Hypothesis**: Same gender tend to be friend with each other

Permutation test on network data

- Observation
  - There are 100 male, 100 female, 4000 edges, 3000 of which occur between the same gender
- Null hypothesis
  - The high number of same-gender relationship occur by chance
- Test statistics
  - Number of same-gender relationship
- Distribution of test statistic under null hypothesis
  - Estimate by generating 100,000 random friendship networks with the same number of people and the same number of interactions
- P-value
  - Fraction of randomized networks with 3000 or more same-gender interactions

Correction for multiple testing

Multiple testing correction

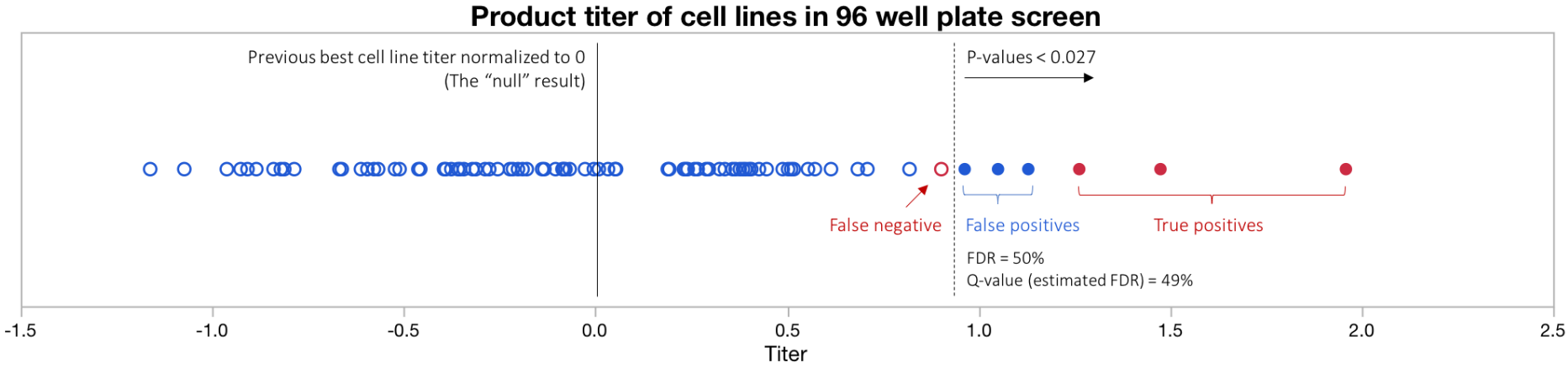
- What is the meaning of setting p-value threshold at 0.05?
  - 5% probability of observing a similar or more extreme result under the null hypothesis
- What if we apply the same p-value threshold to 100 tests?
  - Expected 5% x 100 = 5 tests whose results occurred by chance under the null hypothesis
- This is not acceptable if our conclusion is based on rejecting all null hypotheses

Bonferroni correction

- Divide target p-value cutoff by the number of tests
  - P-value threshold = 0.05
  - Number of test = 100
  - Adjusted p-value cutoff = 0.05 / 100 = 0.0005

- What is the effect?

False discovery rate



- FDR = # of false positive / # of predicted positive
- Directly control the number of false results among tests
  - But calculation of FDR falls under the alternate hypothesis which are often difficult to calculate
- There are ways to control FDR through p-value

Benjamini-Hochberg procedure

- Valid for independent tests (and other nuances)
- Control FDR threshold through p-value
- To aim for FDR of 0.05 from p-values, p_1, p_2, …
  - Sort p-values from small to large
  - Find the smallest *k* such that $\mathbf{p_k \leq 0.05 * k / 10{,}000}$
  - Reject null hypothesis for these *k* smallest p-values

- Extended to Benjamini-Yekutieli procedure
  - Applicable to more dependency between tests

$$P_{(k)} \leq \frac{k}{m \cdot c(m)} \alpha$$

- If the tests are independent or positively correlated (as in Benjamini–Hochberg procedure): $c(m) = 1$
- Under arbitrary dependence (including the case of negative correlation), c(m) is the harmonic number: $c(m) = \sum_{i=1}^{m} \frac{1}{i}$.

Note that $c(m)$ can be approximated by using the Taylor series expansion and the Euler–Mascheroni constant ($\gamma = 0.57721...$):

$$\sum_{i=1}^{m} \frac{1}{i} \approx \ln(m) + \gamma + \frac{1}{2m}.$$

Extra: Bayesian Statistics

Motivation of Bayesian statistics

- Maximum Likelihood Estimator
  - $\boldsymbol{\theta}_{\text{MLE}} = \text{argmax } P(X \mid \boldsymbol{\theta})$
  - What's wrong with this framework?

- We want $\boldsymbol{\theta}_{\text{MAP}} = \text{argmax } P(\boldsymbol{\theta} \mid X)$
  - Maximum a Posteriori (MAP)
  - The most likely $\boldsymbol{\theta}$ given the observed data
  - But do not know how to compute this

- Bayes' rule to the rescue
  - $P(\boldsymbol{\theta} \mid X) = P(X \mid \boldsymbol{\theta}) \times P(\boldsymbol{\theta}) / P(X)$
  - $P(X)$ is constant in this context, $P(\boldsymbol{\theta})$ is called prior
  - $\boldsymbol{\theta}_{\text{MAP}} = \text{argmax } P(X \mid \boldsymbol{\theta}) \times P(\boldsymbol{\theta})$
  - What if we have no prior information on $\boldsymbol{\theta}$?

Maximum A Posteriori (MAP)

- Recall MLE, $\boldsymbol{\theta}_{\text{MLE}} = \text{argmax } P(X \mid \boldsymbol{\theta})$

- But we want $\boldsymbol{\theta}_{\text{MAP}} = \text{argmax } P(\boldsymbol{\theta} \mid X)$
  - The most likely $\boldsymbol{\theta}$ given the observed data
  - We do not know how to compute this

- Bayes' rule to the rescue
  - $P(\theta \mid X) = P(X \mid \theta) \times P(\theta) / P(X)$
  - $P(X)$ is constant, $P(\boldsymbol{\theta})$ is the prior belief on $\boldsymbol{\theta}$
  - $\boldsymbol{\theta}_{\text{MAP}} = \text{argmax } P(X \mid \boldsymbol{\theta}) \times P(\boldsymbol{\theta})$!

Where do prior information come from?

- Provided by domain expert or prior studies
- Estimated from the data
- Noninformative prior
  - Normal distribution with infinite variance
- Conjugate prior
  - Preserve probability family between prior and posterior
  - $P(\boldsymbol{\theta} \mid X) \propto P(X \mid \boldsymbol{\theta}) \times P(\boldsymbol{\theta})$
  - Algebraic convenience

Any question?