BASIC PROBABILITY AND STATISTICS

Saurabh Sinha

Dept. of Biomedical Engineering,
Dept. of Industrial and Systems Engineering,
Georgia Institute of Technology
saurabh.sinha@bme.gatech.edu
https://sites.google.com/view/sinhalaboratorygatech

July 24, 2023

WHY PROBABILITY AND STATISTICS?

- The most mature language we have for understanding data

WHY PROBABILITY AND STATISTICS?

- The most mature language we have for understanding data
- Machine Learning helps with finding patterns in data

Why probability and statistics?

- The most mature language we have for understanding data
- Machine Learning helps with finding patterns in data
- Statistics helps us assess if those patterns are interesting

Why probability and statistics?

- The most mature language we have for understanding data
- Machine Learning helps with finding patterns in data
- Statistics helps us assess if those patterns are interesting
- Machine Learning is often used to make “predictions"

Why probability and statistics?

- The most mature language we have for understanding data
- Machine Learning helps with finding patterns in data
- Statistics helps us assess if those patterns are interesting
- Machine Learning is often used to make “predictions"
- Probability theory can also be used to make those predictions, with confidence estimates

Part I

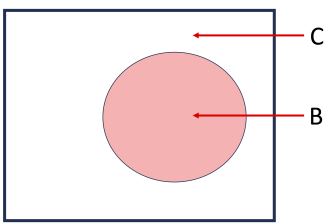
BASICS

Some Basic Rules

- If B and C are mutually exclusive and exhaustive, then $P(A) = P(A, B) + P(A, C)$

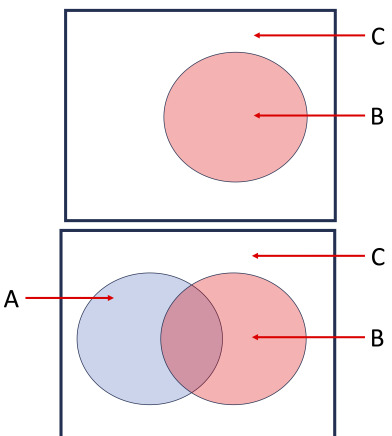
Some Basic Rules

- If B and C are mutually exclusive and exhaustive, then $P(A) = P(A, B) + P(A, C)$



Some Basic Rules

- If B and C are mutually exclusive and exhaustive, then $P(A) = P(A, B) + P(A, C)$



Some Basic Rules

- **Law of Total Probability**: $Pr(A) = \sum_{k=1}^{\infty} Pr(A \cap B_k)$, where the B_j's are exhaustive (union is the entire sample space) and mutually exclusive (no two of them overlap)

Some Basic Rules

- **Law of Total Probability**: $Pr(A) = \sum_{k=1}^{\infty} Pr(A \cap B_k)$, where the B_j's are exhaustive (union is the entire sample space) and mutually exclusive (no two of them overlap)
- **Independent events**: A and B are independent if and only if $P(A, B) = P(A)P(B)$

Some Basic Rules

- **Law of Total Probability**: $Pr(A) = \sum_{k=1}^{\infty} Pr(A \cap B_k)$, where the B_j's are exhaustive (union is the entire sample space) and mutually exclusive (no two of them overlap)
- **Independent events**: A and B are independent if and only if $P(A, B) = P(A)P(B)$
- **Chain rule**: More generally (regardless of independence of A, B), $P(A, B) = P(A)P(B|A) = P(B)P(A|B)$, where $P(X|Y)$ is the conditional probability of X given (conditional on) Y.

Some Basic Rules

- **Law of Total Probability**: $Pr(A) = \sum_{k=1}^{\infty} Pr(A \cap B_k)$, where the B_j's are exhaustive (union is the entire sample space) and mutually exclusive (no two of them overlap)
- **Independent events**: A and B are independent if and only if $P(A, B) = P(A)P(B)$
- **Chain rule**: More generally (regardless of independence of A, B), $P(A, B) = P(A)P(B|A) = P(B)P(A|B)$, where $P(X|Y)$ is the conditional probability of X given (conditional on) Y.
- **Point to remember**: Know when probabilities are added and when they are multiplied.

BAYES' RULE

- Recall $P(A, B) = P(A)P(B|A) = P(B)P(A|B)$.

BAYES' RULE

- Recall $P(A, B) = P(A)P(B|A) = P(B)P(A|B)$.
- Thus, $P(B|A) = \frac{P(A|B)P(B)}{P(A)}$

BAYES' RULE

- Recall $P(A, B) = P(A)P(B|A) = P(B)P(A|B)$.
- Thus, $P(B|A) = \frac{P(A|B)P(B)}{P(A)}$
- This simple relationship is incredibly powerful!

BAYES' RULE

- Recall $P(A, B) = P(A)P(B|A) = P(B)P(A|B)$.
- Thus, $P(B|A) = \frac{P(A|B)P(B)}{P(A)}$
- This simple relationship is incredibly powerful!
- Setting $A = Data$ and $B = Model$ it gives us

$$P(Model|Data) = \frac{P(Data|Model)P(Model)}{P(Data)}$$

WHY DO WE CARE ABOUT BAYES' RULE?

$$P(Model|Data) = \frac{P(Data|Model)P(Model)}{P(Data)}$$

- We often know how to calculate $P(Data|Model)$

WHY DO WE CARE ABOUT BAYES' RULE?

$$P(Model|Data) = \frac{P(Data|Model)P(Model)}{P(Data)}$$

- We often know how to calculate $P(Data|Model)$
- Bayes' rule offers a simple prescription to go from $P(Data|Model)$ to $P(Model|Data)$.

WHY DO WE CARE ABOUT BAYES' RULE?

$$P(Model|Data) = \frac{P(Data|Model)P(Model)}{P(Data)}$$

- We often know how to calculate $P(Data|Model)$
- Bayes' rule offers a simple prescription to go from $P(Data|Model)$ to $P(Model|Data)$.
- $P(Model|Data)$ is good, as it allows us to assess different models (understanding of data) and make predictions about future data.

WHY DO WE CARE ABOUT BAYES' RULE?

$$P(Model|Data) = \frac{P(Data|Model)P(Model)}{P(Data)}$$

- We often know how to calculate $P(Data|Model)$
- Bayes' rule offers a simple prescription to go from $P(Data|Model)$ to $P(Model|Data)$.
- $P(Model|Data)$ is good, as it allows us to assess different models (understanding of data) and make predictions about future data.
- More on Bayesian inference in other lectures, but a highly recommended reading on this: "The Theory That Would Not Die" by Sharon Bertsch McGrayne.

Random Variables

- Random variables are mappings from the sample space Ω to the space R of real numbers.

Random Variables

- Random variables are mappings from the sample space Ω to the space R of real numbers.
- For instance, if we toss a coin, the observation is heads or tails, and we can map these two possible outcomes to the numbers 0 and 1 respectively.

Random Variables

- Random variables are mappings from the sample space Ω to the space R of real numbers.
- For instance, if we toss a coin, the observation is heads or tails, and we can map these two possible outcomes to the numbers 0 and 1 respectively.
- As another example, if we toss a coin say 10 times, then the observation is some sequence of 10 heads and tails, and we can map that observation to a single number, which is the number of heads seen.

Random Variables

- Random variables are mappings from the sample space Ω to the space R of real numbers.
- For instance, if we toss a coin, the observation is heads or tails, and we can map these two possible outcomes to the numbers 0 and 1 respectively.
- As another example, if we toss a coin say 10 times, then the observation is some sequence of 10 heads and tails, and we can map that observation to a single number, which is the number of heads seen.
- Discrete random variable X takes values $x_0, x_1, \ldots x_k \ldots$ with corresponding probabilities $p_0, p_1, \ldots p_k \ldots$, with the condition

$$\sum_{k=0}^{\infty} p_k = 1$$

.

Random Variables

- Random variables are mappings from the sample space Ω to the space R of real numbers.
- For instance, if we toss a coin, the observation is heads or tails, and we can map these two possible outcomes to the numbers 0 and 1 respectively.
- As another example, if we toss a coin say 10 times, then the observation is some sequence of 10 heads and tails, and we can map that observation to a single number, which is the number of heads seen.
- Discrete random variable X takes values $x_0, x_1, \ldots x_k \ldots$ with corresponding probabilities $p_0, p_1, \ldots p_k \ldots$, with the condition

$$\sum_{k=0}^{\infty} p_k = 1$$

.

- The *distribution* of X.

EXPECTATION

- Expectation of X is $E[X] = \sum_k p_k x_k$. Also called “mean".

EXPECTATION

- Expectation of X is $E[X] = \sum_k p_k x_k$. Also called “mean".
- Generalizing, expectation of a function $g(x)$ is $E[g(X)] = \sum_{k=0}^{\infty} p_k g(x_k)$

EXPECTATION

- Expectation of X is $E[X] = \sum_k p_k x_k$. Also called “mean".
- Generalizing, expectation of a function $g(x)$ is $E[g(X)] = \sum_{k=0}^{\infty} p_k g(x_k)$
- Expectation is a very intuitive thing: the common “average".

EXPECTATION

- Expectation of X is $E[X] = \sum_k p_k x_k$. Also called “mean".
- Generalizing, expectation of a function $g(x)$ is $E[g(X)] = \sum_{k=0}^{\infty} p_k g(x_k)$
- Expectation is a very intuitive thing: the common “average".
- Let X = number of heads when you toss a fair coin 10 times. X is a random variable with possible values $x_0 = 0, x_1 = 1, ..., x_{10} = 10$, with respective probabilities $p_0, p_1, ..., p_{10}$ that you can calculate with a formula and maybe a calculator.

EXPECTATION

- Expectation of X is $E[X] = \sum_k p_k x_k$. Also called “mean".
- Generalizing, expectation of a function $g(x)$ is $E[g(X)] = \sum_{k=0}^{\infty} p_k g(x_k)$
- Expectation is a very intuitive thing: the common “average".
- Let X = number of heads when you toss a fair coin 10 times. X is a random variable with possible values $x_0 = 0, x_1 = 1, ..., x_{10} = 10$, with respective probabilities $p_0, p_1, ..., p_{10}$ that you can calculate with a formula and maybe a calculator.
- The expectation of X is far easier to calculate: you intuitively know it's $E(X) = 5$, without a calculator.

EXPECTATION

- Expectation of X is $E[X] = \sum_k p_k x_k$. Also called “mean".
- Generalizing, expectation of a function $g(x)$ is $E[g(X)] = \sum_{k=0}^{\infty} p_k g(x_k)$
- Expectation is a very intuitive thing: the common “average".
- Let X = number of heads when you toss a fair coin 10 times. X is a random variable with possible values $x_0 = 0, x_1 = 1, ..., x_{10} = 10$, with respective probabilities $p_0, p_1, ..., p_{10}$ that you can calculate with a formula and maybe a calculator.
- The expectation of X is far easier to calculate: you intuitively know it's $E(X) = 5$, without a calculator.
- Linearity of expectation: $E(X + Y) = E(X) + E(Y)$. This makes many expectation calculations easy!

VARIANCE

- Variance: $E[(X - E(X))^2]$

VARIANCE

- Variance: $E[(X - E(X))^2]$
- In plain English, it's the average “deviation" from the the mean. (“deviation" $\equiv$ square of difference)

VARIANCE

- Variance: $E[(X - E(X))^2]$
- In plain English, it's the average “deviation" from the the mean. (“deviation" $\equiv$ square of difference)
- Standard Deviation is the positive square root of variance.

VARIANCE

- Variance: $E[(X - E(X))^2]$
- In plain English, it's the average “deviation" from the the mean. (“deviation" $\equiv$ square of difference)
- Standard Deviation is the positive square root of variance.
- Often, beginners compare means of two groups and make claims. For instance, “my Machine Learning program has average accuracy of 80% compared to this other program whose average accuracy is 75%." You can't make these claims without also looking into the variance.

VARIANCE

- Variance: $E[(X - E(X))^2]$
- In plain English, it's the average "deviation" from the the mean. ("deviation" $\equiv$ square of difference)
- Standard Deviation is the positive square root of variance.
- Often, beginners compare means of two groups and make claims. For instance, "my Machine Learning program has average accuracy of 80% compared to this other program whose average accuracy is 75%." You can't make these claims without also looking into the variance.
- Variance can be harder to calculate analytically. For instance, the following is NOT TRUE in general: $Var(X + Y) = Var(X) + Var(Y)$.

Discrete Distributions I: Binomial distribution

- Bernoulli trial: an experiment with two possible outcomes: e.g., “success” with prob. p and “failure” with prob. $1 - p$.

Discrete Distributions I: Binomial distribution

- Bernoulli trial: an experiment with two possible outcomes: e.g., “success” with prob. p and “failure” with prob. $1 - p$.
- Suppose you do N independent Bernoulli trials, each with same success prob. What is the distribution of the number of successes?

Discrete Distributions I: Binomial distribution

- Bernoulli trial: an experiment with two possible outcomes: e.g., "success" with prob. p and "failure" with prob. $1 - p$.
- Suppose you do N independent Bernoulli trials, each with same success prob. What is the distribution of the number of successes?
- Binomial distribution: prob. of k successes in N trials.

$$p_k = \binom{N}{k} p^k (1-p)^{N-k}$$

Discrete Distributions I: Binomial distribution

- Bernoulli trial: an experiment with two possible outcomes: e.g., "success" with prob. p and "failure" with prob. $1 - p$.
- Suppose you do N independent Bernoulli trials, each with same success prob. What is the distribution of the number of successes?
- Binomial distribution: prob. of k successes in N trials.

$$p_k = \binom{N}{k} p^k (1-p)^{N-k}$$

- Mean $E(X) = Np$; Variance $\mathrm{Var}(X) = Np(1-p)$

Discrete Distributions I: Binomial distribution

- Bernoulli trial: an experiment with two possible outcomes: e.g., "success" with prob. p and "failure" with prob. $1 - p$.
- Suppose you do N independent Bernoulli trials, each with same success prob. What is the distribution of the number of successes?
- Binomial distribution: prob. of k successes in N trials.

$$p_k = \binom{N}{k} p^k (1-p)^{N-k}$$

- Mean $E(X) = Np$; Variance $\text{Var}(X) = Np(1-p)$
- Example: A genome has 20% 'C's, 20% 'G's, 30% 'A's, 30% 'T's. Find all 1 Kbp segments with G/C content that is at least three standard deviations above expectation. (Answer: Expectation $= 1000 \times 0.4 = 400$, Standard deviation $= \sqrt{1000 * 0.4 * 0.6} \approx 15$, so look for all 1 Kbp segments with G+C count above 445.)

Discrete Distributions II: Poisson distribution

- Used to model the number of occurrences of “events" over a fixed interval of time, e.g., number of 911 calls in an hour.

$$p_k = \frac{e^{-\lambda}\lambda^k}{k!}$$

Discrete Distributions II: Poisson distribution

- Used to model the number of occurrences of “events" over a fixed interval of time, e.g., number of 911 calls in an hour.

$$p_k = \frac{e^{-\lambda}\lambda^k}{k!}$$

- λ is a parameter.

Discrete Distributions II: Poisson distribution

- Used to model the number of occurrences of “events" over a fixed interval of time, e.g., number of 911 calls in an hour.

$$p_k = \frac{e^{-\lambda}\lambda^k}{k!}$$

- λ is a parameter.
- $E(X) = \lambda$, $\text{Var}(X) = \lambda$

Discrete Distributions II: Poisson distribution

- Used to model the number of occurrences of “events" over a fixed interval of time, e.g., number of 911 calls in an hour.

$$p_k = \frac{e^{-\lambda}\lambda^k}{k!}$$

- λ is a parameter.
- $E(X) = \lambda$, $\text{Var}(X) = \lambda$
- Example: The pattern “TCACGT" has about one occurrence per 1000 bp in a genome. What is the probability of observing four or more occurrences of “TCACGT" in a particular gene’s promoter (1000 bp long)? (Answer: $\sim 2\%$)

Continuous Distributions

- With continuous distributions, probability of any particular value is 0. We talk about "probability density" at a particular value, not its probability.

Continuous Distributions

- With continuous distributions, probability of any particular value is 0. We talk about "probability density" at a particular value, not its probability.
- Probability density function (pdf) $f(x)$ and cumulative prob. distribution $F(x) = P[X \leq x]$.

$$F(x) = \int_{-\infty}^{x} f(x)dx$$

Continuous Distributions

- With continuous distributions, probability of any particular value is 0. We talk about "probability density" at a particular value, not its probability.
- Probability density function (pdf) $f(x)$ and cumulative prob. distribution $F(x) = P[X \leq x]$.

$$F(x) = \int_{-\infty}^{x} f(x)dx$$

- $E[x] = \int xf(x)dx$

Continuous Distributions

- With continuous distributions, probability of any particular value is 0. We talk about "probability density" at a particular value, not its probability.
- Probability density function (pdf) $f(x)$ and cumulative prob. distribution $F(x) = P[X \leq x]$.

$$F(x) = \int_{-\infty}^{x} f(x)dx$$

- $E[x] = \int xf(x)dx$
- $Var[x] = \int (x - E[X])^2 f(x)dx$

Normal distribution

- Normal distribution

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right]$$

- μ and σ are parameters equal to expectation and standard deviation resp.
- Supported on the whole set of reals
- The famous Bell curve!

Part II

STATISTICAL TESTING

Statistical Test: informally speaking

- Toss a coin 50 times, get heads 50 times. I ask you "Is the coin a fair coin?"

Statistical Test: informally speaking

- Toss a coin 50 times, get heads 50 times. I ask you "Is the coin a fair coin?"
- You do not want to believe the "hypothesis" that “the coin is unbiased”. i.e., you reject this hypothesis.

Statistical Test: informally speaking

- Toss a coin 50 times, get heads 50 times. I ask you "Is the coin a fair coin?"
- You do not want to believe the "hypothesis" that “the coin is unbiased”. i.e., you reject this hypothesis.
- Same experiment, but 40 heads and 10 tails. Do you still want to reject the hypothesis?

Statistical Test: informally speaking

- Toss a coin 50 times, get heads 50 times. I ask you "Is the coin a fair coin?"
- You do not want to believe the "hypothesis" that “the coin is unbiased”. i.e., you reject this hypothesis.
- Same experiment, but 40 heads and 10 tails. Do you still want to reject the hypothesis?
- Maybe? Maybe not? Need a systematic procedure.

Statistical Test: informally speaking

- Toss a coin 50 times, get heads 50 times. I ask you "Is the coin a fair coin?"
- You do not want to believe the "hypothesis" that “the coin is unbiased”. i.e., you reject this hypothesis.
- Same experiment, but 40 heads and 10 tails. Do you still want to reject the hypothesis?
- Maybe? Maybe not? Need a systematic procedure.
- That’s what a statistical test does.

STATISTICAL TESTING: A NOT-SO-GOOD SOLUTION

- Step 1: State the null hypothesis. “The coin is unbiased". This allows probability calculations. For instance, coin tosses are Bernoulli trials with prob. of heads $p = 0.5$.

STATISTICAL TESTING: A NOT-SO-GOOD SOLUTION

- Step 1: State the null hypothesis. “The coin is unbiased". This allows probability calculations. For instance, coin tosses are Bernoulli trials with prob. of heads $p = 0.5$.
- Step 2: Decide upon the “test statistic", a random variable whose value depends on the data. For instance, X = “number of heads in 50 coin tosses".

STATISTICAL TESTING: A NOT-SO-GOOD SOLUTION

- Step 1: State the null hypothesis. “The coin is unbiased". This allows probability calculations. For instance, coin tosses are Bernoulli trials with prob. of heads $p = 0.5$.
- Step 2: Decide upon the “test statistic", a random variable whose value depends on the data. For instance, X = “number of heads in 50 coin tosses".
- Step 3. Calculate the probability of the observed value of test statistic (X). For instance, $P(X = 40) = 0.000009$. (Use Binomial distribution.)

STATISTICAL TESTING: A NOT-SO-GOOD SOLUTION

- Step 1: State the null hypothesis. “The coin is unbiased". This allows probability calculations. For instance, coin tosses are Bernoulli trials with prob. of heads $p = 0.5$.
- Step 2: Decide upon the “test statistic", a random variable whose value depends on the data. For instance, X = “number of heads in 50 coin tosses".
- Step 3. Calculate the probability of the observed value of test statistic (X). For instance, $P(X = 40) = 0.000009$. (Use Binomial distribution.)
- Step 4. If the probability from Step 3 is “small" (say less than 5%), then reject the null hypothesis – the data do not seem likely under that hypothesis, so it is likely to be wrong.

WHERE'S THE PROBLEM?

- Say you saw 30 heads out of 50 coin tosses. What does your intuition say? Is it very unlikely?

WHERE'S THE PROBLEM?

- Say you saw 30 heads out of 50 coin tosses. What does your intuition say? Is it very unlikely?
- No. You probably are not surprised to see 30 heads in 50 coin tosses, when using a regular (fair) coin.

WHERE'S THE PROBLEM?

- Say you saw 30 heads out of 50 coin tosses. What does your intuition say? Is it very unlikely?
- No. You probably are not surprised to see 30 heads in 50 coin tosses, when using a regular (fair) coin.
- Yet, Step 3 will calculate the probability $P(X = 30) = 0.042$ and in Step 4 you will reject the null hypothesis.

WHERE'S THE PROBLEM?

- Say you saw 30 heads out of 50 coin tosses. What does your intuition say? Is it very unlikely?
- No. You probably are not surprised to see 30 heads in 50 coin tosses, when using a regular (fair) coin.
- Yet, Step 3 will calculate the probability $P(X = 30) = 0.042$ and in Step 4 you will reject the null hypothesis.
- This problem gets exacerbated as you have more data!

WHERE'S THE PROBLEM?

- Say you saw 30 heads out of 50 coin tosses. What does your intuition say? Is it very unlikely?
- No. You probably are not surprised to see 30 heads in 50 coin tosses, when using a regular (fair) coin.
- Yet, Step 3 will calculate the probability $P(X = 30) = 0.042$ and in Step 4 you will reject the null hypothesis.
- This problem gets exacerbated as you have more data!
- Say you saw 250 heads out of 500 coin tosses. Surely this is not unlikely under the fair coin hypothesis? It's the most likely outcome after all! Yet, Step 3 calculates $P(X = 250) = 0.036$ and you will reject the null hypothesis!

WHERE'S THE PROBLEM?

- Say you saw 30 heads out of 50 coin tosses. What does your intuition say? Is it very unlikely?
- No. You probably are not surprised to see 30 heads in 50 coin tosses, when using a regular (fair) coin.
- Yet, Step 3 will calculate the probability $P(X = 30) = 0.042$ and in Step 4 you will reject the null hypothesis.
- This problem gets exacerbated as you have more data!
- Say you saw 250 heads out of 500 coin tosses. Surely this is not unlikely under the fair coin hypothesis? It's the most likely outcome after all! Yet, Step 3 calculates $P(X = 250) = 0.036$ and you will reject the null hypothesis!
- Our test is not so good!

STATISTICAL TESTING: THE COMMON SOLUTION

- Step 1: State the null hypothesis. "The coin is unbiased". (Coin tosses are Bernoulli trials with prob. of heads $p = 0.5$.)

STATISTICAL TESTING: THE COMMON SOLUTION

- Step 1: State the null hypothesis. "The coin is unbiased". (Coin tosses are Bernoulli trials with prob. of heads $p = 0.5$.)
- Step 2: Decide upon the test statistic, X = number of heads in 50 coin tosses.

STATISTICAL TESTING: THE COMMON SOLUTION

- Step 1: State the null hypothesis. "The coin is unbiased". (Coin tosses are Bernoulli trials with prob. of heads $p = 0.5$.)
- Step 2: Decide upon the test statistic, X = number of heads in 50 coin tosses.
- Step 3. Calculate the probability of the observed value of test statistic (X) *being equal to or more "extreme" than the observed value*. For instance, $P(X \geq 40) = 0.0000119$. (Use Binomial distribution.)

STATISTICAL TESTING: THE COMMON SOLUTION

- Step 1: State the null hypothesis. "The coin is unbiased". (Coin tosses are Bernoulli trials with prob. of heads $p = 0.5$.)
- Step 2: Decide upon the test statistic, X = number of heads in 50 coin tosses.
- Step 3. Calculate the probability of the observed value of test statistic (X) *being equal to or more "extreme" than the observed value*. For instance, $P(X \geq 40) = 0.0000119$. (Use Binomial distribution.)
- Step 4. If the probability from Step 3 is "small" (say less than 5%), then reject the null hypothesis – the data do not seem likely under that hypothesis, so it is likely to be wrong.

STATISTICAL TESTING: THE COMMON SOLUTION

- Step 1: State the null hypothesis. "The coin is unbiased". (Coin tosses are Bernoulli trials with prob. of heads $p = 0.5$.)
- Step 2: Decide upon the test statistic, X = number of heads in 50 coin tosses.
- Step 3. Calculate the probability of the observed value of test statistic (X) *being equal to or more "extreme" than the observed value*. For instance, $P(X \geq 40) = 0.0000119$. (Use Binomial distribution.)
- Step 4. If the probability from Step 3 is "small" (say less than 5%), then reject the null hypothesis – the data do not seem likely under that hypothesis, so it is likely to be wrong.
- Note: The "5%" threshold defining "small" in Step 4 is called the "significance level" of the test.

STATISTICAL TESTING: THE COMMON SOLUTION

- Step 1: State the null hypothesis. "The coin is unbiased". (Coin tosses are Bernoulli trials with prob. of heads $p = 0.5$.)
- Step 2: Decide upon the test statistic, X = number of heads in 50 coin tosses.
- Step 3. Calculate the probability of the observed value of test statistic (X) *being equal to or more "extreme" than the observed value*. For instance, $P(X \geq 40) = 0.0000119$. (Use Binomial distribution.)
- Step 4. If the probability from Step 3 is "small" (say less than 5%), then reject the null hypothesis – the data do not seem likely under that hypothesis, so it is likely to be wrong.
- Note: The "5%" threshold defining "small" in Step 4 is called the "significance level" of the test.
- Note: The probability calculated in Step 3 is called the "p-value" of the test.

POINTS TO PONDER (ON YOUR OWN)

- Let's see if the "problem" got fixed. First, 30 heads out of 50 tosses gives us $P(X \geq 30) = 0.10$, which is not that small. Second, 250 heads out of 500 coin tosses gives us $P(X \geq 250) = 0.52$, clearly not a small number.

POINTS TO PONDER (ON YOUR OWN)

- Let's see if the "problem" got fixed. First, 30 heads out of 50 tosses gives us $P(X \geq 30) = 0.10$, which is not that small. Second, 250 heads out of 500 coin tosses gives us $P(X \geq 250) = 0.52$, clearly not a small number.
- Another point: what does "more extreme" mean, in the language of Step 3?

POINTS TO PONDER (ON YOUR OWN)

- Let's see if the "problem" got fixed. First, 30 heads out of 50 tosses gives us $P(X \geq 30) = 0.10$, which is not that small. Second, 250 heads out of 500 coin tosses gives us $P(X \geq 250) = 0.52$, clearly not a small number.
- Another point: what does "more extreme" mean, in the language of Step 3?
- When you see 40 heads out of 50 tosses, you're really testing if it's "too many", so "as extreme" means $X \geq 40$.

POINTS TO PONDER (ON YOUR OWN)

- Let's see if the "problem" got fixed. First, 30 heads out of 50 tosses gives us $P(X \geq 30) = 0.10$, which is not that small. Second, 250 heads out of 500 coin tosses gives us $P(X \geq 250) = 0.52$, clearly not a small number.
- Another point: what does "more extreme" mean, in the language of Step 3?
- When you see 40 heads out of 50 tosses, you're really testing if it's "too many", so "as extreme" means $X \geq 40$.
- You might also be testing 40 out of 50 is "too far from what you expected (25)", in which case you should calculate $P(X \geq 40) + P(X \leq 10)$, since both $X = 40$ and $X = 10$ are as far from your expectation and thus "as extreme".

POINTS TO PONDER (ON YOUR OWN)

- Let's see if the "problem" got fixed. First, 30 heads out of 50 tosses gives us $P(X \geq 30) = 0.10$, which is not that small. Second, 250 heads out of 500 coin tosses gives us $P(X \geq 250) = 0.52$, clearly not a small number.
- Another point: what does "more extreme" mean, in the language of Step 3?
- When you see 40 heads out of 50 tosses, you're really testing if it's "too many", so "as extreme" means $X \geq 40$.
- You might also be testing 40 out of 50 is "too far from what you expected (25)", in which case you should calculate $P(X \geq 40) + P(X \leq 10)$, since both $X = 40$ and $X = 10$ are as far from your expectation and thus "as extreme".
- In the former scenario, we say "Null hypothesis: $p = 0.5$, Alternative hypothesis: $p > 0.5$. A "one-sided test".

POINTS TO PONDER (ON YOUR OWN)

- Let's see if the "problem" got fixed. First, 30 heads out of 50 tosses gives us $P(X \geq 30) = 0.10$, which is not that small. Second, 250 heads out of 500 coin tosses gives us $P(X \geq 250) = 0.52$, clearly not a small number.
- Another point: what does "more extreme" mean, in the language of Step 3?
- When you see 40 heads out of 50 tosses, you're really testing if it's "too many", so "as extreme" means $X \geq 40$.
- You might also be testing 40 out of 50 is "too far from what you expected (25)", in which case you should calculate $P(X \geq 40) + P(X \leq 10)$, since both $X = 40$ and $X = 10$ are as far from your expectation and thus "as extreme".
- In the former scenario, we say "Null hypothesis: $p = 0.5$, Alternative hypothesis: $p > 0.5$. A "one-sided test".
- In the latter scenario, we say "Null hypothesis: $p = 0.5$, Alternative hypothesis: $p \neq 0.5$. A "two-sided test".

NULL DISTRIBUTION OF P-VALUE

- Let's say that the null hypothesis (unbiased coin) is true. Do 50 coin tosses, count heads. This count is a random variable, call it X.

NULL DISTRIBUTION OF P-VALUE

- Let's say that the null hypothesis (unbiased coin) is true. Do 50 coin tosses, count heads. This count is a random variable, call it X.
- Calculate the p-value of X, e.g., with a one-sided test. Denote the p-value by p. Note that p is determined by X, and is thus a random variable itself.

NULL DISTRIBUTION OF P-VALUE

- Let's say that the null hypothesis (unbiased coin) is true. Do 50 coin tosses, count heads. This count is a random variable, call it X.
- Calculate the p-value of X, e.g., with a one-sided test. Denote the p-value by p. Note that p is determined by X, and is thus a random variable itself.
- What is the probability distribution of p?

NULL DISTRIBUTION OF P-VALUE

- Let's say that the null hypothesis (unbiased coin) is true. Do 50 coin tosses, count heads. This count is a random variable, call it X.
- Calculate the p-value of X, e.g., with a one-sided test. Denote the p-value by p. Note that p is determined by X, and is thus a random variable itself.
- What is the probability distribution of p?
- $\Pr(p \leq \pi) = \pi$.

NULL DISTRIBUTION OF P-VALUE

- Let's say that the null hypothesis (unbiased coin) is true. Do 50 coin tosses, count heads. This count is a random variable, call it X.
- Calculate the p-value of X, e.g., with a one-sided test. Denote the p-value by p. Note that p is determined by X, and is thus a random variable itself.
- What is the probability distribution of p?
- $\Pr(p \leq \pi) = \pi$.
- In other words, the p-value follows a uniform distribution between 0 and 1.

NULL DISTRIBUTION OF P-VALUE

- Let's say that the null hypothesis (unbiased coin) is true. Do 50 coin tosses, count heads. This count is a random variable, call it X.
- Calculate the p-value of X, e.g., with a one-sided test. Denote the p-value by p. Note that p is determined by X, and is thus a random variable itself.
- What is the probability distribution of p?
- $\Pr(p \leq \pi) = \pi$.
- In other words, the p-value follows a uniform distribution between 0 and 1.
- This means that *even if an unbiased coin was used* there's a 5% chance that your test will produce a "significant" p-value of ≤ 0.05, and you will reject the null hypothesis (that coin is unbiased), i.e., you'll make an erroneous inference.

NULL DISTRIBUTION OF P-VALUE

- Let's say that the null hypothesis (unbiased coin) is true. Do 50 coin tosses, count heads. This count is a random variable, call it X.
- Calculate the p-value of X, e.g., with a one-sided test. Denote the p-value by p. Note that p is determined by X, and is thus a random variable itself.
- What is the probability distribution of p?
- $\Pr(p \leq \pi) = \pi$.
- In other words, the p-value follows a uniform distribution between 0 and 1.
- This means that *even if an unbiased coin was used* there's a 5% chance that your test will produce a "significant" p-value of ≤ 0.05, and you will reject the null hypothesis (that coin is unbiased), i.e., you'll make an erroneous inference.
- This is an important realization. We'll come back to it later.

STATISTICAL TEST EXAMPLE: T-TEST

- Consider two groups of hypertension patients, each of size K. The first group (“M") is given a medication, while the second group (“P") was given placebo. Measure blood pressure in each individual.

STATISTICAL TEST EXAMPLE: T-TEST

- Consider two groups of hypertension patients, each of size K. The first group (“M") is given a medication, while the second group (“P") was given placebo. Measure blood pressure in each individual.
- Assume that blood pressure in both groups is a normally distributed variable (X_M and X_P). Null hypothesis: X_M and X_P have the same mean and variance (i.e., all measurements in either group are from the same probability distribution).

STATISTICAL TEST EXAMPLE: T-TEST

- Calculate averages $\overline{X_M}$ and $\overline{X_P}$ respectively and the standard deviations s_M and s_P respectively for both groups.

STATISTICAL TEST EXAMPLE: T-TEST

- Calculate averages $\overline{X_M}$ and $\overline{X_P}$ respectively and the standard deviations s_M and s_P respectively for both groups.
- Calculate the "statistic"

$$t = \frac{\overline{X_M} - \overline{X_P}}{\sqrt{\frac{1}{K}(s_M^2 + s_P^2)}}$$

STATISTICAL TEST EXAMPLE: T-TEST

- Calculate averages $\overline{X_M}$ and $\overline{X_P}$ respectively and the standard deviations s_M and s_P respectively for both groups.
- Calculate the "statistic"

$$t = \frac{\overline{X_M} - \overline{X_P}}{\sqrt{\frac{1}{K}(s_M^2 + s_P^2)}}$$

- The statistic t follows a "t-distribution with $2(K-1)$ degrees of freedom". This allows determining p-value of an observed value of t.

STATISTICAL TEST EXAMPLE: T-TEST

- Calculate averages $\overline{X_M}$ and $\overline{X_P}$ respectively and the standard deviations s_M and s_P respectively for both groups.
- Calculate the "statistic"

$$t = \frac{\overline{X_M} - \overline{X_P}}{\sqrt{\frac{1}{K}(s_M^2 + s_P^2)}}$$

- The statistic t follows a "t-distribution with $2(K-1)$ degrees of freedom". This allows determining p-value of an observed value of t.
- This is the "t-test".

STATISTICAL TEST EXAMPLE: T-TEST

- Calculate averages $\overline{X_M}$ and $\overline{X_P}$ respectively and the standard deviations s_M and s_P respectively for both groups.
- Calculate the "statistic"

$$t = \frac{\overline{X_M} - \overline{X_P}}{\sqrt{\frac{1}{K}(s_M^2 + s_P^2)}}$$

- The statistic t follows a "t-distribution with $2(K-1)$ degrees of freedom". This allows determining p-value of an observed value of t.
- This is the "t-test".
- Example use: you could compare a gene's expression in two groups of biospecimens (e.g., patients and healthy subjects) using the t-test, to determine if this gene is of interest. (We'll come back to this in a bit.)

ENRICHMENT TEST OR HYPERGEOMETRIC TEST

- A study is looking at a “population” of N genes.

Enrichment test or Hypergeometric test

- A study is looking at a “population” of N genes.
- A subset of n genes have been identified as being turned on in cancer.

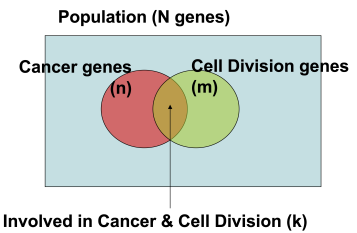
Enrichment test or Hypergeometric test

- A study is looking at a “population” of N genes.
- A subset of n genes have been identified as being turned on in cancer.
- Suspiciously many of these n cancer genes are known to be involved in “cell division”.

Enrichment test or Hypergeometric test

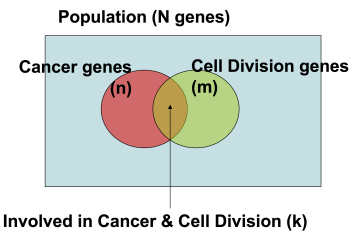
- A study is looking at a “population” of N genes.
- A subset of n genes have been identified as being turned on in cancer.
- Suspiciously many of these n cancer genes are known to be involved in “cell division”.
- Can we demonstrate a connection?

ENRICHMENT TEST OR HYPERGEOMETRIC TEST



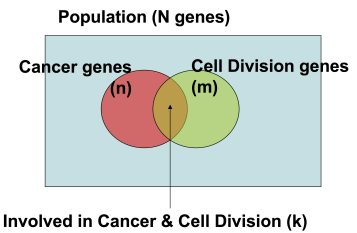
- A study is looking a "population" of N genes.
- A subset of n genes have been identified as being turned on in cancer.
- Collect the set of all genes involved in cell division, say this is of size m

ENRICHMENT TEST OR HYPERGEOMETRIC TEST



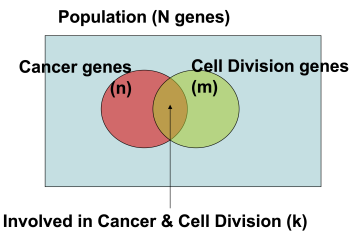
- A study is looking a "population" of N genes.
- A subset of n genes have been identified as being turned on in cancer.
- Collect the set of all genes involved in cell division, say this is of size m
- Find k genes to be in the intersection of the cancer set and the cell division set

ENRICHMENT TEST OR HYPERGEOMETRIC TEST



- A study is looking a "population" of N genes.
- A subset of n genes have been identified as being turned on in cancer.
- Collect the set of all genes involved in cell division, say this is of size m
- Find k genes to be in the intersection of the cancer set and the cell division set
- Is this (k) significantly large, given N, m, n?

ENRICHMENT TEST OR HYPERGEOMETRIC TEST



- A study is looking a "population" of N genes.
- A subset of n genes have been identified as being turned on in cancer.
- Collect the set of all genes involved in cell division, say this is of size m
- Find k genes to be in the intersection of the cancer set and the cell division set
- Is this (k) significantly large, given N, m, n?

Enrichment test or Hypergeometric test

- The Hypergeometric test:

Enrichment test or Hypergeometric test

- The Hypergeometric test:
- Let us keep the n cancer genes to be a fixed set.

Enrichment test or Hypergeometric test

- The Hypergeometric test:
- Let us keep the n cancer genes to be a fixed set.
- If we had picked a random sample of m genes, how likely is an intersection equal to k?

Enrichment test or Hypergeometric test

- The Hypergeometric test:
- Let us keep the n cancer genes to be a fixed set.
- If we had picked a random sample of m genes, how likely is an intersection equal to k?
- $f(k; N, n, m) = \frac{\binom{n}{k}\binom{N-n}{m-k}}{\binom{N}{m}}$

Enrichment test or Hypergeometric test

- The Hypergeometric test:
- Let us keep the n cancer genes to be a fixed set.
- If we had picked a random sample of m genes, how likely is an intersection equal to k?
- $f(k; N, n, m) = \frac{\binom{n}{k}\binom{N-n}{m-k}}{\binom{N}{m}}$
- How likely is an intersection equal to or greater than k?

Enrichment test or Hypergeometric test

- The Hypergeometric test:
- Let us keep the n cancer genes to be a fixed set.
- If we had picked a random sample of m genes, how likely is an intersection equal to k?
- $f(k; N, n, m) = \frac{\binom{n}{k}\binom{N-n}{m-k}}{\binom{N}{m}}$
- How likely is an intersection equal to or greater than k?
- $P = \sum_{j \geq k} f(j; N, n, m)$

Enrichment test or Hypergeometric test

- The Hypergeometric test:
- Let us keep the n cancer genes to be a fixed set.
- If we had picked a random sample of m genes, how likely is an intersection equal to k?
- $f(k; N, n, m) = \frac{\binom{n}{k}\binom{N-n}{m-k}}{\binom{N}{m}}$
- How likely is an intersection equal to or greater than k?
- $P = \sum_{j \geq k} f(j; N, n, m)$
- If P calculated this way is below some threshold α, e.g., 0.05, we say that the association between the cancer set and the cell division set is statistically significant.

ENRICHMENT TEST OR HYPERGEOMETRIC TEST

- The Hypergeometric test:
- Let us keep the n cancer genes to be a fixed set.
- If we had picked a random sample of m genes, how likely is an intersection equal to k?
- $f(k; N, n, m) = \frac{\binom{n}{k}\binom{N-n}{m-k}}{\binom{N}{m}}$
- How likely is an intersection equal to or greater than k?
- $P = \sum_{j \geq k} f(j; N, n, m)$
- If P calculated this way is below some threshold α, e.g., 0.05, we say that the association between the cancer set and the cell division set is statistically significant.
- In other words, we have just discovered a link between cancer and cell division, which is probably worthy of further investigation.

TESTING A GENE FOR DIFFERENTIAL EXPRESSION

- Suppose a gene's expression was measured in 100 different samples from cancer patients and 100 samples from healthy individuals

TESTING A GENE FOR DIFFERENTIAL EXPRESSION

- Suppose a gene's expression was measured in 100 different samples from cancer patients and 100 samples from healthy individuals
- Test whether gene is "differentially expressed" between the two groups: t-test.

TESTING A GENE FOR DIFFERENTIAL EXPRESSION

- Suppose a gene's expression was measured in 100 different samples from cancer patients and 100 samples from healthy individuals
- Test whether gene is "differentially expressed" between the two groups: t-test.
- Test produces a p-value, and if this p-value is $\leq \alpha$ (say $\alpha = 0.05$), we can proclaim this gene to be "differentially expressed" in cancer. Interesting!

FALSE POSITIVES?

- We noted previously that even if the null hypothesis is true, i.e., the gene is not significantly different between cancer patients and healthy individuals, the test may call it “differentially expressed" and interesting. The probability of such a “false positive" prediction is α.

FALSE POSITIVES?

- We noted previously that even if the null hypothesis is true, i.e., the gene is not significantly different between cancer patients and healthy individuals, the test may call it “differentially expressed" and interesting. The probability of such a “false positive" prediction is α.
- False Positive: “Positive" because rejecting null hypothesis usually implicates the gene as being interesting in some way. “False" because null hypothesis being true means that the rejection was an error.

FALSE POSITIVES?

- We noted previously that even if the null hypothesis is true, i.e., the gene is not significantly different between cancer patients and healthy individuals, the test may call it “differentially expressed" and interesting. The probability of such a “false positive" prediction is α.
- False Positive: “Positive" because rejecting null hypothesis usually implicates the gene as being interesting in some way. “False" because null hypothesis being true means that the rejection was an error.
- So yes, our statistical test can make a false positive error, but such errors are “controlled" (probability of the error is α).

TESTING MULTIPLE GENES

- Now consider repeating the above test on 10000 genes, one by one.

TESTING MULTIPLE GENES

- Now consider repeating the above test on 10000 genes, one by one.
- In each test, the probability of false positive is α.

TESTING MULTIPLE GENES

- Now consider repeating the above test on 10000 genes, one by one.
- In each test, the probability of false positive is α.
- In other words, if I do 10000 tests, I might make false positive errors $10000 \times \alpha$ times. (For $\alpha = 0.05$, this amounts to 500 false predictions!)

TESTING MULTIPLE GENES

- Now consider repeating the above test on 10000 genes, one by one.
- In each test, the probability of false positive is α.
- In other words, if I do 10000 tests, I might make false positive errors $10000 \times \alpha$ times. (For $\alpha = 0.05$, this amounts to 500 false predictions!)
- This is the multiple hypothesis testing problem. A significance level (α) that looks convincing on a single test no longer looks so convincing when doing many tests.

TESTING MULTIPLE GENES

- Now consider repeating the above test on 10000 genes, one by one.
- In each test, the probability of false positive is α.
- In other words, if I do 10000 tests, I might make false positive errors $10000 \times \alpha$ times. (For $\alpha = 0.05$, this amounts to 500 false predictions!)
- This is the multiple hypothesis testing problem. A significance level (α) that looks convincing on a single test no longer looks so convincing when doing many tests.
- We'd like to predict a set of genes as being interesting, i.e., as violating null hypothesis, but with "control" over the total false positive error.

TESTING MULTIPLE GENES

- One option: make the significance level α *really* small.

TESTING MULTIPLE GENES

- One option: make the significance level α *really* small.
- For example, when testing 10000 genes, if $\alpha = 0.05/10000$, the same calculation tells us that the expected number of false positive errors is $10000 \times \alpha = 0.05$. This is quite acceptable. It's called the Bonferroni correction.

TESTING MULTIPLE GENES

- One option: make the significance level α *really* small.
- For example, when testing 10000 genes, if $\alpha = 0.05/10000$, the same calculation tells us that the expected number of false positive errors is $10000 \times \alpha = 0.05$. This is quite acceptable. It's called the Bonferroni correction.
- The Bonferroni correction is harsh! If we demand that p-value is $\leq \alpha = 0.05/10000$, very few genes may show up as significant. Did we overdo this "multiple testing correction" thing? We went from using $\alpha = 0.05$ to $\alpha = 0.000005$. Is there some "middle ground" here, that allows us to keep false positive errors low without resorting to such a super small α?

TESTING MULTIPLE GENES

- One option: make the significance level α *really* small.
- For example, when testing 10000 genes, if $\alpha = 0.05/10000$, the same calculation tells us that the expected number of false positive errors is $10000 \times \alpha = 0.05$. This is quite acceptable. It's called the Bonferroni correction.
- The Bonferroni correction is harsh! If we demand that p-value is $\leq \alpha = 0.05/10000$, very few genes may show up as significant. Did we overdo this "multiple testing correction" thing? We went from using $\alpha = 0.05$ to $\alpha = 0.000005$. Is there some "middle ground" here, that allows us to keep false positive errors low without resorting to such a super small α?
- FDR ("false discovery rate") is one such middle ground.

False Discovery Rate (self-reading)

- Is a procedure for deciding upon a significance level so that the false positive errors are controlled at a low level.

False Discovery Rate (self-reading)

- Is a procedure for deciding upon a significance level so that the false positive errors are controlled at a low level.
- The final outcome will be a set of genes predicted to be differentially expressed

False Discovery Rate (self-reading)

- Is a procedure for deciding upon a significance level so that the false positive errors are controlled at a low level.
- The final outcome will be a set of genes predicted to be differentially expressed
- We will have some control on the proportion of false positives among these predicted genes

False Discovery Rate (self-reading)

- Is a procedure for deciding upon a significance level so that the false positive errors are controlled at a low level.
- The final outcome will be a set of genes predicted to be differentially expressed
- We will have some control on the proportion of false positives among these predicted genes
- The theory talks about 'tests' and not 'genes', of course. Here, we are using it in the context of tests involving genes.

False Discovery Rate (self-reading)

- Say there are 10000 genes and 100 are truly differentially expressed. It is probably OK then to predict some set of 100 genes as being differentially expressed, with the disclaimer that (say) 10 of these may be false positives.

False Discovery Rate (self-reading)

- Say there are 10000 genes and 100 are truly differentially expressed. It is probably OK then to predict some set of 100 genes as being differentially expressed, with the disclaimer that (say) 10 of these may be false positives.
- Our testing is based on p-values. So we need a way to go from a p-value (e.g., “probability of a false positive call on this gene is 0.05") to an overall false positive proportion (e.g., “of all genes found significant by us, we expect 10% to be false positives").

An FDR Procedure (self-reading)

- Proposed by Benjamini and Hochberg in 1995. Many other procedures since then, but we'll only see this original one.
- Begin with a per-gene p-value, i.e., $Pr(X \geq \tau | H_0)$, for every one of the g genes being studied.
- Let the g p-values be denoted by $p_{(i)}$
- Consider these g p-values to be sorted in ascending order, i.e., $p_{(1)} \leq p_{(2)} \leq \ldots \leq p_{(g)}$
- Let $H_0^{(i)}$ be the null hypothesis corresponding to $p(i)$
- Let $q_i = \frac{i\alpha}{g}$ for $i = 1, 2, 3 \ldots g$ where α is the desired FDR
- Let k be the max i such that $p_{(i)} \leq q_i$
- Procedure: Reject null hypothesis $H_0^{(1)}, H_0^{(2)}, \ldots H_0^{(k)}$ and accept all others.
- Theorem: This controls the FDR at level α. *What does that mean?*

A NOTE ON FDRS VS P-VALUES (SELF-READING)

- FDR is fundamentally different from a p-value.
- P-value assesses significance of data. If we publish some data that we claim to be significant, we should present a small p-value for the data (e.g., ≤ 0.05)
- FDR is generally used as a “culling tool"; the investigator wants to predict a set of genes to test experimentally, and an FDR of 0.1 or even 0.5 may be acceptable to them (they will do twice as much experimental work, which may be fine)