Math 481

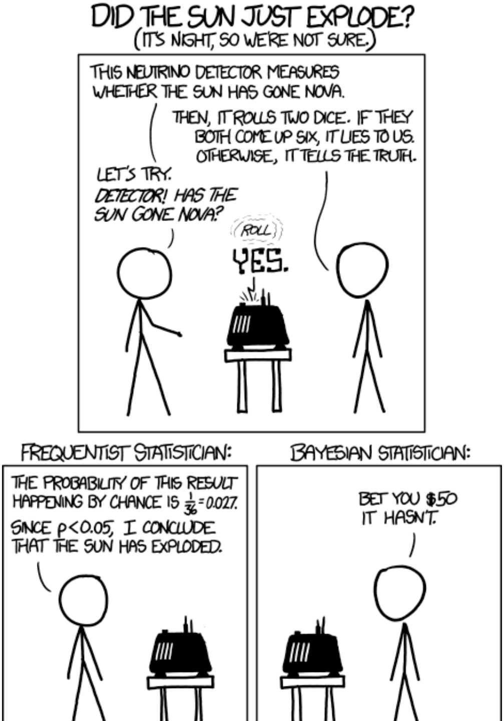
- Instructor: Chenxi Wu wuchenxi2013@gmail.com
- Office: Hill 434, Office hours: 10-11 am Tu, Wed or by appointment, starting from Jan 28.
- Grading policy: 10% weekly homework (lowest dropped), 20% each of the two midterms, 50% final exam.
- Prerequisite: Probability. Will finish review of basic probability on Feb 12.
- Weekly assignments: 2-3 homework problems a week, grade for correctness, similar to exams. There will also be questions from textbook assigned for practice which you don't need to hand in.
- No late homework or make up midterms.

Main topics we will cover:

- Review of probability
- Point estimate
- p-values and hypothesis testing
- Confidence intervals
- Bayesian statistics

Bayesian and non-Bayesian approaches to statistics

- Non-Bayesian approach: Set up a null hypothesis and try to show that observation is highly unlikely if null hypothesis is true.
- Bayesian approach: Assume prior distribution of some parameter, calculate posterior via Bayes formula


DID THE SUN JUST EXPLODE?
(IT'S NIGHT, SO WE'RE NOT SURE.)
THIS NEUTRINO DETECTOR MEASURES WHETHER THE SUN HAS GONE NOVA.
THEN, IT ROLLS TWO DICE. IF THEY BOTH COME UP SIX, IT LIES TO US. OTHERWISE, IT TELLS THE TRUTH.
LET'S TRY. DETECTOR! HAS THE SUN GONE NOVA?
ROLL
YES.
FREQUENTIST STATISTICIAN:
THE PROBABILITY OF THIS RESULT HAPPENING BY CHANCE IS $\frac{1}{36}=0.027$. SINCE $p<0.05$, I CONCLUDE THAT THE SUN HAS EXPLODED.
BAYESIAN STATISTICIAN:
BET YOU $50 IT HASN'T.

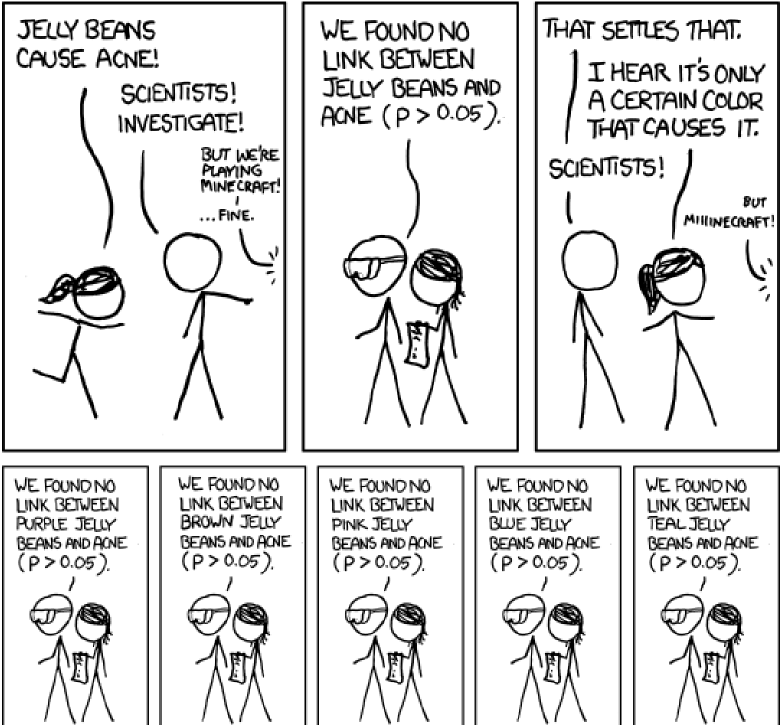
Some review of basic probability

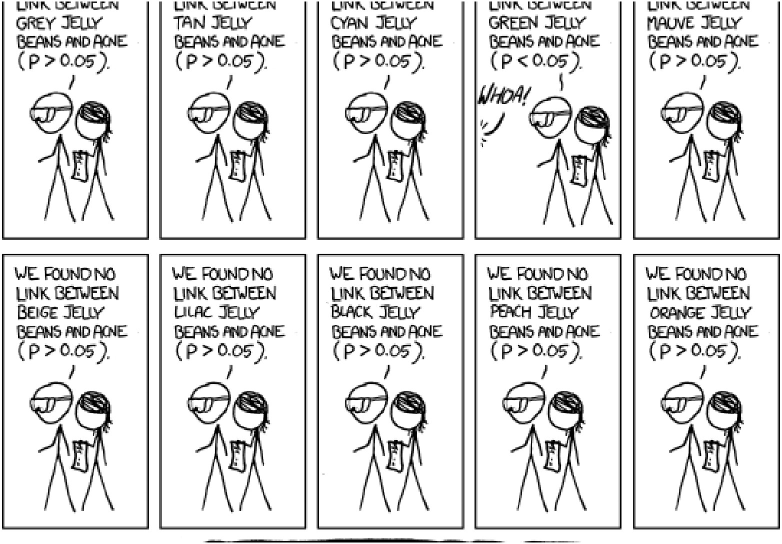
- Two random events A and B are called **independent** if $P(A \cap B) = P(A)P(B)$
- If A and B are two random events, $P(A) > 0$. The conditional probability of B when A is given is $P(B|A) = P(A \cap B)/P(A)$.

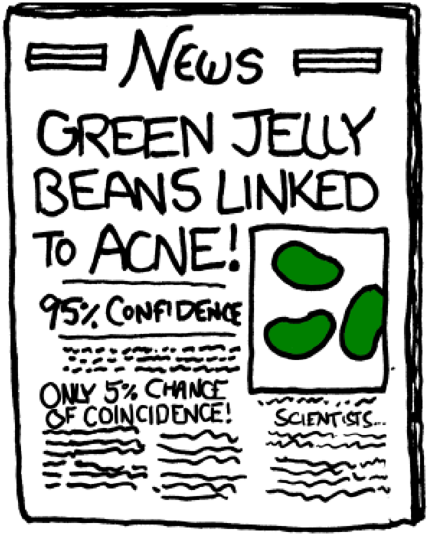
Example

Suppose you are given a coin, you flip it 5 times and get head on all 5 of them.

- Suppose the coin is fair, what is the odds that it gets head for 5 times in 5 flips?
- **Null hypothesis**
- **p-value**


JELLY BEANS CAUSE ACNE!
SCIENTISTS! INVESTIGATE!
BUT WE'RE PLAYING MINECRAFT!
... FINE.
WE FOUND NO LINK BETWEEN JELLY BEANS AND ACNE (P > 0.05).
THAT SETTLES THAT.
I HEAR IT'S ONLY A CERTAIN COLOR THAT CAUSES IT.
SCIENTISTS!
BUT MIIIINECRAFT!
WE FOUND NO LINK BETWEEN PURPLE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BROWN JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN PINK JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BLUE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN TEAL JELLY BEANS AND ACNE (P > 0.05).


GREY JELLY BEANS AND ACNE (P > 0.05).
TAN JELLY BEANS AND ACNE (P > 0.05).
CYAN JELLY BEANS AND ACNE (P > 0.05).
GREEN JELLY BEANS AND ACNE (P < 0.05).
WHOA!
MAUVE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BEIGE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN LILAC JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BLACK JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN PEACH JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN ORANGE JELLY BEANS AND ACNE (P > 0.05).


News
GREEN JELLY BEANS LINKED TO ACNE!
95% CONFIDENCE
ONLY 5% CHANCE OF COINCIDENCE!
SCIENTISTS...

- Suppose the coin is biased and gets head at probability p.
  - What is the probability that it gets head for 5 times in 5 flips?
  - What is the p that maximizes this probability?
  - What is the range of p such that the probability for 5 heads in 5 flips is no less than 0.05?
- **Maximum likelihood estimate (MLE)**
- **Confidence interval**

- Suppose you pick the coin among a pile of 100 coins, 99 of which is fair and 1 has head on both sides. What is the chance of the coin being unfair given the results of the 5 flips?
- **Prior and posterior**

- Suppose the odds for getting a head is uniformly distributed in $[0, 1]$, given the results of the 5 flips, what do you think is the most likely value for p? How about the expectation?
- **Maximum a posteriori (MAP) estimate**

Basic definitions in probability

A **Probability** is a triple (S, F, P) where S is called the **sample space** denoting all possible states of the world, $F \subset \mathcal{P}(S)$ the **event space** and $P : F \rightarrow \mathbb{R}$ a real-valued function on F, such that:

1. F is closed under complement and countable union.
2. P is non negative.
3. $P(S) = 1$
4. If $\{E_i\}$ is a countable sequence of disjoint events in F, $P(\bigcup_i E_i) = \sum_i P(E_i)$.

Random variables

- A (real valued) **random variable** X is a function $S \to \mathbb{R}$ such that the preimage of any open interval is in F. Multivariant random variables can be defined similarly.
- The **cumulative distribution function (cdf)** of a random variable X is $F(x) = P(X \leq x)$.
- If $F(x) = \int_{-\infty}^{x} f(t)dt$ we call f the **probability density function (pdf)**
- If there is a countable set C and $g : C \to \mathbb{R}$ such that $F(x) = \sum_{y \in C, y \leq x} g(y)$ we call X **discrete** and g the **probability distribution**
- The **expectation** of a random variable X is defined as $E[X] = \int_S XdP$.

For those who know analysis

- A probability is a measure $P : F \to \mathbb{R}$, where F is a σ-algebra on sample space S and $P(S) = 1$.
- A random variable X is a P-measurable function on S.
- The expectation of a random variable X is the integral $\int_S X dP$.

Some questions

- Must the cdf of a random variable be left or right continuous?
- X is the number of heads in 2 fair coin flips. What is the cdf of X? What is the expectation of X? What is the expectation of $(X - E[X])^2$?
- Can you write down a random variable that is neither discrete nor has a pdf?
- Can you write down a random variable which has no expectation?

Independence and conditional probability

- X and Y are 2 random variables, X and Y are independent iff $F_{X,Y}(s,t) = P(X \leq s \cap Y \leq t) = F_X(s)F_Y(t)$.
- If A is some event with non zero probability, $F_{X|A}(s) = P(X \leq s|A) = P(X \leq s \cap A)/P(A)$.
- If X and Y has joint p.d.f. $f_{X,Y}$ with non zero marginal density f_Y, then $f_{X|Y=a}(s) = f_{X,Y}(s,a)/f_Y(a)$.
- If A_i are disjoint events with non zero probabilities, $B \subset \mathbb{R}$, $P(X \in B| \cup_i A_i) = \sum_i (P(A_i)P(X \in B|A_i))/\sum_i P(A_i)$.
- If Y has p.d.f. f_Y, $A \subset \mathbb{R}$such that $P(Y \in A) > 0$, B is a random event, then $P(B|Y \in A) = \int_A f_Y(s)P(B|Y = s)ds/P(Y \in A)$.

Special random variables

- **Discrete**: Takes on countably values, has p.d.
- **Continuous**: has p.d.f.

2 random variables X and Y has the same distribution iff they have the same c.d.f., or for any $A \subset \mathbb{R}$, $P(X \in A) = P(Y \in A)$. Random variables with the same distribution are NOT necessarily the same.

Special Probability distributions

- **Bernoulli distribution**: $f(1) = \theta$, $f(0) = 1 - \theta$.
- **Binomial distribution** (sum of iid Bernoulli):
  $f(x) = \binom{n}{x} \theta^x (1-\theta)^{n-x}$, $x = 0, 1, \ldots, n$.
- **Negative Binomial distribution** (waiting time for the k-th success of iid trials): $f(x) = \binom{x-1}{k-1} \theta^k (1-\theta)^{x-k}$, $x = k, k+1, \ldots$. When $k = 1$ it is the **geometric distribution**.
- **Hypergeometric distribution** (randomly pick n elements at random from N elements, the number of elements picked from a fixed subset of M elements)
  $f(x) = \binom{M}{x} \binom{N-M}{n-x} \binom{N}{n}^{-1}$.

- **Poisson distribution** (limit of binomial as $n \to \infty$, $n\theta \to \lambda$) $f(x) = \lambda^x e^{-\lambda}/x!$.
- **Multinomial distribution** $f(x_1, \ldots x_k) = \left(\begin{array}{c} n \\ x_1, \ldots, x_k \end{array} \right) \theta_1^{x_1} \ldots \theta_k^{x_k}$, $\sum_i x_i = n$, $\theta_i \theta_i = 1$.
- **Multivariate Hypergeometric distribution** $f(x_1, \ldots, x_k) = \prod_i \left(\begin{array}{c} M_i \\ x_i \end{array} \right) \cdot \left(\begin{array}{c} N \\ n \end{array} \right)^{-1}$. $\sum_i x_i = n$, $\sum_i M_i = N$.

Special Probability Density Functions

$\Gamma(a) = \int_0^\infty x^{a-1} e^{-x} dx$. $\Gamma(k) = (k-1)!$ when $k = 1, 2, \ldots$.

- **Uniform distribution**: $f(x) = \begin{cases} 1/(b-a) & x \in (a,b) \\ 0 & x \notin (a,b) \end{cases}$.
- **Normal distribution**: $f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$.
- **Multivariate Normal distribution**: $x \in \mathbb{R}^d$, Σ positive definite $d \times d$ symmetric matrix, $f(x) = (2\pi)^{-d/2} |\Sigma|^{-1/2} e^{-\frac{1}{2}(x-\mu)^T \Sigma^{-1} (x-\mu)}$.
- χ^2 **distribution** d: degrees of freedom. Squared sum of d normal distributions: $f(x) = \begin{cases} \frac{1}{2^{d/2}\Gamma(d/2)} x^{\frac{d-2}{2}} e^{-x/2} & x > 0 \\ 0 & x \leq 0 \end{cases}$.

- **Exponential distribution** $f(x) = \begin{cases} \frac{1}{\theta} e^{-x/\theta} & x > 0 \\ 0 & x \le 0 \end{cases}$.
- **Gamma-distribution**: $f(x) = \begin{cases} \frac{1}{\beta^\alpha \Gamma(\alpha)} x^{\alpha-1} e^{-x/\beta} & x > 0 \\ 0 & x \le 0 \end{cases}$
- **Beta distribution**: (conjugate prior of Bernoulli distribution) $f(x) = \begin{cases} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1} & x \in (0,1) \\ 0 & x \notin (0,1) \end{cases}$.

Example: If the bias of a coin p has a uniform **prior** in $[0,1]$, after n flips there are a heads and b tails, the **posterior** will be Beta distribution with $\alpha = a+1$, $\beta = b+1$.

Sample mean and sample variance

X_i i.i.d. (independent with identical distribution)

- **Sample mean**: $\overline{X} = \frac{1}{n} \sum_i X_i$
- **Sample variance**:
  $S^2 = \frac{1}{n-1} \sum_i (X_i - \overline{X})^2 = \frac{1}{n-1} (\sum_i X_i^2 - n\overline{X}^2)$.

Properties:

- $E[\overline{X}] = E[X_1]$
- $Var(\overline{X}) = \frac{1}{n} Var(X_1)$
- $\sqrt{\frac{n}{Var(X_1)}} (\overline{X} - E[X_1]) \to \mathcal{N}(0, 1)$ (Central Limit Theorem)
- $E[S^2] = Var(X_1)$

Assuming $X_i \sim \mathcal{N}(\mu, \sigma^2)$:

- $\overline{X}$ and S^2 are independent.
- $\overline{X} \sim \mathcal{N}(\mu, \frac{\sigma^2}{n})$
- $\frac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1)$

Proof of $\frac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1)$

$$(n-1)S^2 = \sum_i (X_i - \overline{X})^2 = \sum_i ((X_i^2 - E[X_i]) - (\overline{X} - E[\overline{X}]))^2$$

$$= \sum_i (X_i^2 - E[X_i])^2 - n(\overline{X} - E[\overline{X}])^2$$

Now divide by σ^2, the first term is $\chi^2(n)$ and second $\chi^2(1)$.

χ^2 distribution

Definition: X_i independent, $\mathcal{N}(0,1)$, then $\sum_{i=1}^{n} X_i = \chi^2(n)$
PDF:

$$f(x) = \begin{cases} \frac{1}{n/2\Gamma(n/2)} x^{\frac{n-2}{2}} e^{-x/2} & x > 0 \\ 0 & x \leq 0 \end{cases}$$

Calculation of PDF:

$$f_{\chi^2(n)}(r) = \frac{d}{dr} \int_{\sum_i x_i^2 \leq r} (2\pi)^{-n/2} e^{-\sum_i x_i^2/2} dx_1 \ldots dx_n$$

$$= (2\pi)^{-n/2} e^{-r/2} \frac{d}{dr} Vol(B(\sqrt{r}))$$

Where $B(x)$ is the ball of radius x.

t distribution

Definition: X and Y independent, $X \sim \mathcal{N}(0,1)$, $Y \sim \chi^2(n)$, then $\frac{X}{\sqrt{Y/n}} \sim t(n)$.
By LLN, when $n \to \infty$ this converges to $\mathcal{N}(0,1)$.
PDF:

$$f(x) = \frac{\Gamma(\frac{n+1}{2})}{\sqrt{\pi n}\Gamma(\frac{n}{2})}(1 + \frac{x^2}{n})^{-\frac{n+1}{2}}$$

Calculation of PDF of t

$$f_{t(n)}(s) = \frac{d}{ds}P(X \leq s\sqrt{Y/n}) = \frac{d}{ds}\int_0^\infty dy \int_{-\infty}^{s\sqrt{y/n}}$$

$$dx \frac{1}{\sqrt{2\pi}} e^{-x^2/2} \frac{1}{2^{n/2}\Gamma(n/2)} y^{\frac{n-2}{2}} e^{-y/2}$$

$$= \int_0^\infty dy \sqrt{y/n} \frac{1}{\sqrt{2\pi}} e^{-s^2 y/2n} \frac{1}{2^{n/2}\Gamma(n/2)} y^{\frac{n-2}{2}} e^{-y/2}$$

$$= \frac{1}{\sqrt{2\pi n} 2^{n/2}\Gamma(n/2)} \int_0^\infty dy y^{\frac{n-1}{2}} e^{-y(1+\frac{s^2}{n})/2}$$

Now let $z = y(1 + \frac{s^2}{n})/2$ and it's done.

F-distribution

Definition: U and V independent, $U \sim \chi^2(m)$, $V \sim \chi^2(n)$, then $\frac{U/m}{V/n} \sim F(m,n)$
CDF:

$$f(x) = \begin{cases} \frac{\Gamma(\frac{m+n}{2})}{\Gamma(\frac{m}{2})\Gamma(\frac{n}{2})}(\frac{m}{n})^{m/2} x^{m/2-1}(1+\frac{m}{n}x)^{-\frac{m+n}{2}} & x > 0 \\ 0 & x \leq 0 \end{cases}$$

Strategy for calculating the PDF of $Y = g(X_i)$:

1. Find joint pdf of X_i
2. Write down the CDF of Y as a probability, hence, some integral of the pdf of X_i
3. Differentiate the CDF of Y.

Probability Review

- Probability, cdf and pdf for continuous random variables:
  - **Probability to cdf**: $F_X(t) = P(X \leq t)$
  - **cdf to pdf**: $f_X(t) = \frac{d}{dt} F_X(t)$
  - **pdf to probability**: $P(X \in A) = \int_A f_X(s)ds$
- Probability, cdf and pd for discrete random variables:
  - **Probability to cdf**: $F_X(t) = P(X \leq t)$
  - **cdf to pd**: $F_X(t) = \sum_{s \leq t} g_X(s)$
  - **pd to probability**: $P(X \in A) = \sum_{s \in A} g_X(s)$
- Joint cdf/pdf/pd, independence, conditional probability.
- Expectation, variance, covariance
- LLN and CLT
- Special distributions: binomial, uniform, normal, χ^2, etc.

Point estimates

Basic setting:

- $\mathcal{F}$: a family of possible distributions (represented by a family of cdf, pdf, or pd)
- $\theta : \mathcal{F} \to \mathbb{R}$ population parameter
- $X_1, \ldots X_n$ i.i.d. with distribution $F \in \mathcal{F}$
- $\hat{\theta} = \hat{\theta}(X_1, \ldots, X_n)$ a function of X_i, which is an estimate of $\theta(F)$, is called a point estimate.

Example: $\mathcal{F}$: all distributions with an expectation, then $\overline{X}$ is a point estimate of the expectation.

$\hat{\theta}$ is a point estimate of θ.

- The **bias** is $E[\hat{\theta}] - \theta$. $\hat{\theta}$ is called unbiased if $E[\hat{\theta}] = \theta$.
- The **variance** is $Var(\hat{\theta})$.
- $\hat{\theta}$ is called **minimum variance unbiased estimate** if it has the smallest variance among all unbiased estimates.
- $\hat{\theta}_1$ and $\hat{theta}_2$ are two unbiased estimates, the relative efficiency is the ratio of their variance. When they are biased, one can use the mean squared error $E[(\hat{\theta} - \theta)^2]$ instead.
- $\hat{\beta}$ is called **asymptotically unbiased** if bias converges to 0 as $n \to \infty$.
- $\hat{\beta}$ is called **consistent** if $\hat{\beta}$ converges to β in distribution.

Review of definitions regarding point estimates

$\hat{\theta}$ is a point estimate of θ

- Unbiased
- Minimal Variance Unbiased
- Asymptotically unbiased
- Consistent

Properties:

- Minimal Variance Unbiased can be verified via Cramer-Rao
- Mean squared error
  $E[(\hat{\theta}-\theta)^2] = E[((\hat{\theta}-E[\hat{\theta}])+(E[\hat{\theta}-\theta))^2] = Var(\hat{\theta})+(E[\hat{\theta}]-\theta)^2$
- Mean squared error $\to 0$ implies consistence:

$$P(|\hat{\theta}-\theta| > \epsilon) < \frac{E[(\hat{\theta}-\theta)^2]}{\epsilon^2}$$

But consistence does not imply mean squared error $\to 0$.

MLE

Suppose $X_i \sim F(\theta)$, i.i.d., observation is $x_1, \ldots, x_k$, then $\hat{\theta} = \arg\max_\theta L(x_1, \ldots x_k, \theta)$.

- When F is a continuous distribution with p.d.f. $f(x, \theta)$, let $L(x_1, \ldots, x_k, \theta) = \prod_i f(x_i, \theta)$
- When F is a discrete distribution with p.d. $g(x, \theta)$, let $L(x_1, \ldots, x_k, \theta) = \prod_i g(x_i, \theta)$