Probability Notes

Jeremy Teitelbaum

Probability Basics

Outcomes and Sample Space

Probability begins with a set X of "outcomes". This set may be continuous or discrete.

- $X = \{H, T\}$, the result of a single coin flip. (discrete)
- X is the possible results of throwing two six-sided dice – ordered pairs. (discrete)
- X is the set of real numbers, where a value x means measuring the temperature $t_0 + x$ where t_0 is the "true" temperature. (continous)

The set X of possible outcomes is called the *sample space.*

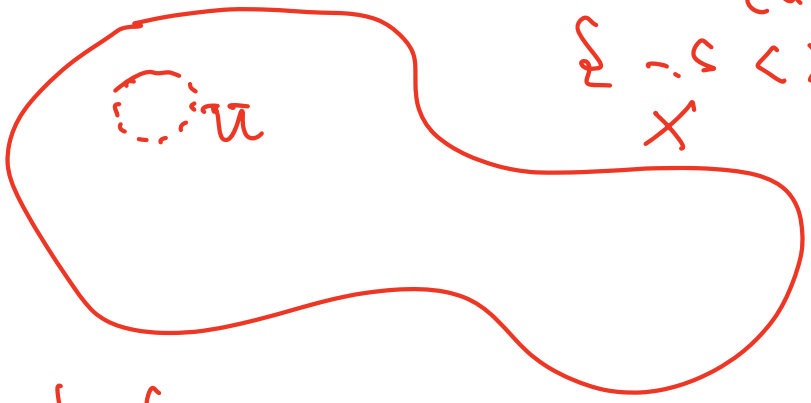
Event

An "event" is a subset of the sample space – a collection of outcomes.

The probability function P takes values between 0 and 1 and measures the "chance" that an event "occurs."

If a sequence of events are disjoint, then the probability of them all happening is the sum of their probabilities.

$$P(U_1 \cup \cdots \cup U_n) = \sum_{i=1}^{n} P(U_i)$$



Discrete case

$X \to$ 5 flips of coin

Sample Space $\hookrightarrow$ { HTTHT } (5 flips)

32 outcomes

Event: you have 2 heads E

$E = \{ \underline{HHTTT}, \ldots \}$

$|E| = 10$

$= \binom{5}{2} = \frac{5!}{2!3!}$

$= \frac{5 \cdot 4}{2}$

$P(HHTTT) = \frac{1}{32}$

$P(E) = P(HHTTT) + P(HTHTT) + \ldots = 10/32$

Events - discrete examples

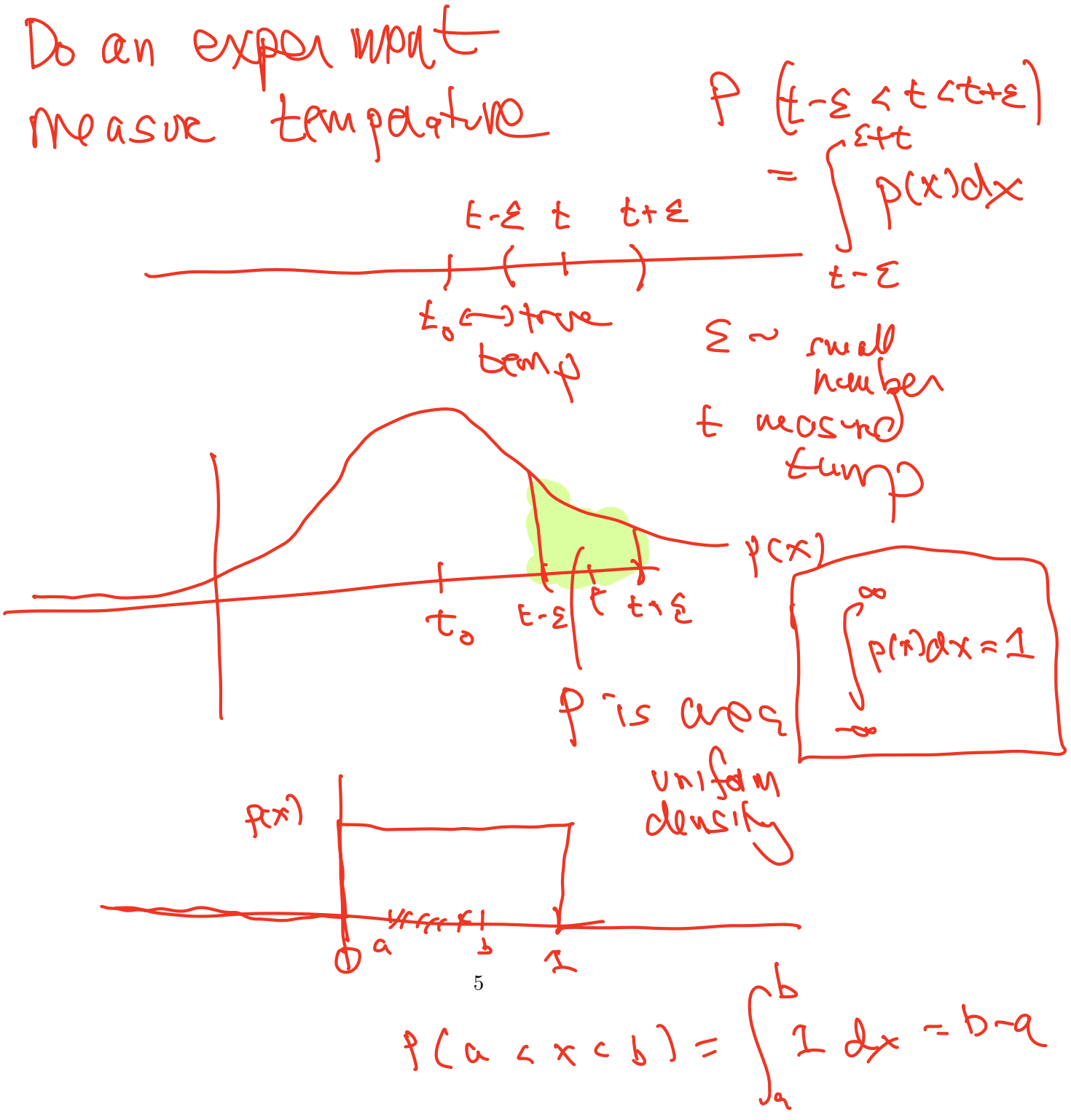
- $P(\{H\}) = 1/2$
- $P(\{(\boxdot, \boxdot)\}) = 1/36$
- the probability of the event E consisting of throwing two dice that sum to 5:

$$E = \{(\boxdot, \boxdot), (\boxdot, \boxdot), (\boxdot, \boxdot), (\boxdot, \boxdot)\}$$

is $(4)(1/36) = 1/9$

Events - continuous example

- $X = \mathbf{R}$.
- Probability arises from a density function $p(x)$
- $P(U) = \int_U p(x)dx$
- $\int_{-\infty}^{\infty} p(x)dx = 1$.



Normal distribution

- Measure temperature t using a thermometer.
- True temperature is t_0.
- Error $x = t - t_0$

$$P(|t - t_0| < \delta) = \int_{x=-\delta}^{\delta} p_\sigma(x)dx$$

where

$$p_\sigma(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-x^2/(2\sigma^2)}.$$

σ is called the "standard deviation".

Normal distribution cont'd

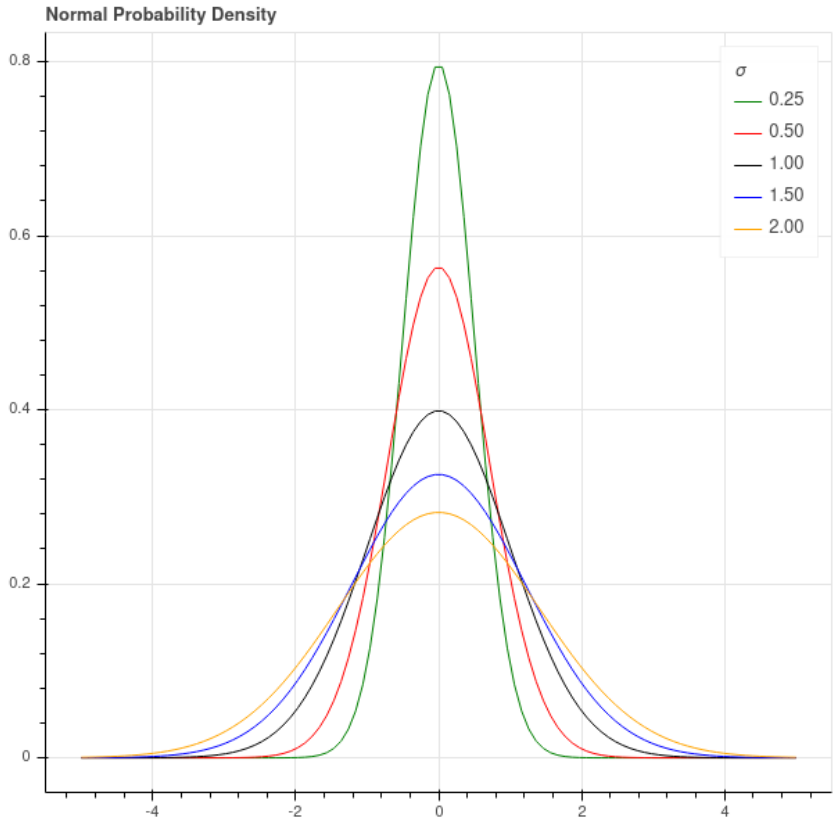


Figure 1: Normal Distributions

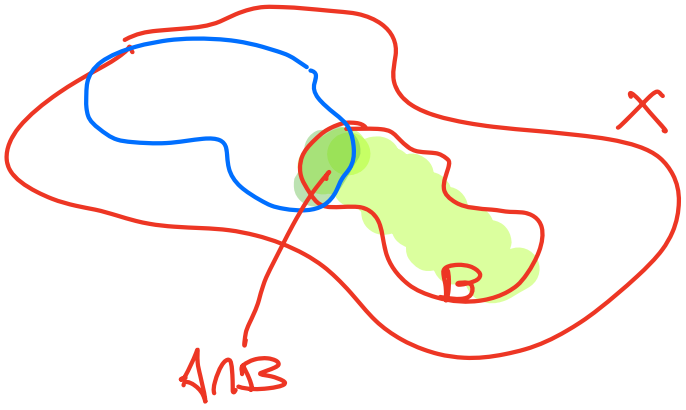
Conditional Probability and Bayes Theorem

Conditional Probability

- Draw a card from a deck. $P(\text{king}) = 4/52 = 1/13$.
- Now suppose you *know* the card is a face card. Given that information, the probability of drawing a king is $4/12 = 1/3$, This is an example of *conditional probability.* $P(A|B)$.
- More generally

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

Meaning the chance of getting A among conditions where B is known to hold.



Bayes Theorem

Theorem: Given events A and B in a sample space X, we have

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

Proof: Substitute $P(B \cap A)/P(A)$ for $P(B|A)$.

$$P(B|A) = \frac{P(A \cap B)}{P(A)}$$

$$P(B|A)P(A) = P(A \cap B)$$

$$\frac{P(B|A)P(A)}{P(B)} = \frac{P(A \cap B)}{P(B)} = P(A|B)$$

An example: COVID testing

- A person can be *sick* (S) or *well* (W).
- Their covid test can be *positive* (+) or *negative* (-).

There are four possibilities:

- S+ – this is a true positive
- S- – this is a false negative
- W+ – this is a false positive
- W- – this is a true negative

An early CDC report estimated that $P(+|W) = 1/200$ and $P(-|S) = 1/4$.

What is $P(S|+)$?

COVID testing continued

- Suppose I get a covid test and it's positive. How likely am I to have the disease? In other words, what is $P(S|+)$?

The answer depends on the prevalence $p = P(S)$, the chance that I have COVID in the first place.

$$P(S|+) = \frac{P(S,+)}{P(+)}$$

$P(+|W) = .005$ $\quad P(-|S) = .25$

$P(-|W) = .995$ $\quad P(+|S) = .75$

$$P(S|+) = \frac{P(+|S)\,P(S)}{P(+)}$$

$P(S) \longleftrightarrow$ population incidence (p)

$$P(+) = P(S,+) + P(W,+)$$

$$= P(+|S)P(S) + P(+|W)P(W) \qquad P(W) = 1-p$$

$$.75p + (.005)(1-p)$$

$$\boxed{P(S|+) = \frac{.750p}{.5 + .745p}}$$

COVID testing continued

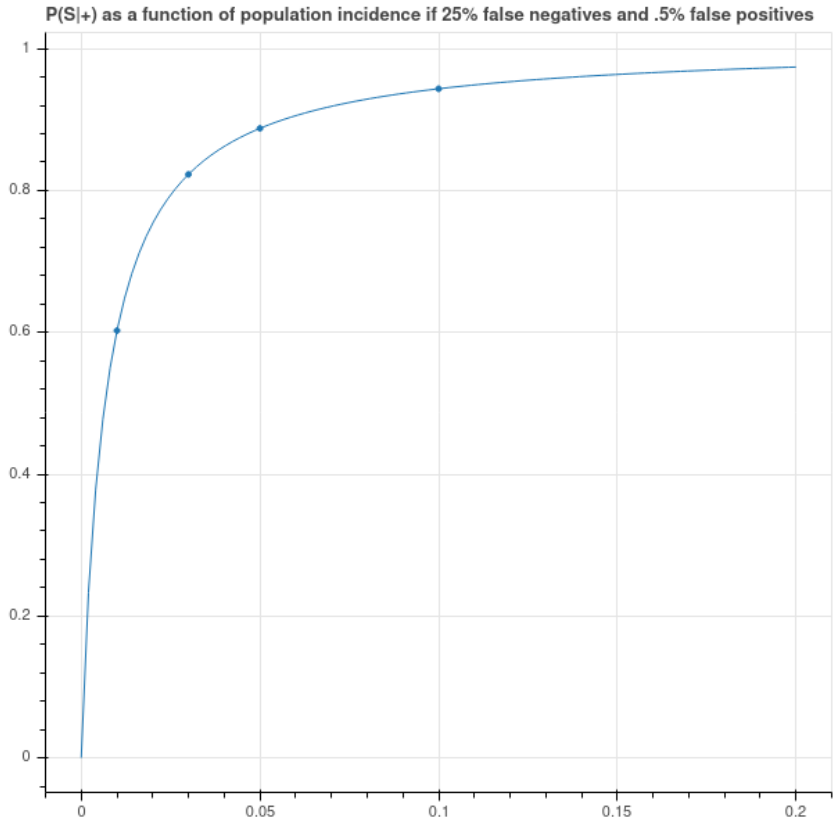


Figure 2: Chance I have COVID if I get a positive test vs prevalence

Independence

Independence

Definition: Two events A and B are independent if $P(A \cap B) = P(A)P(B)$. Alternatively, they are independent if $P(A|B) = P(A)$ and $P(B|A) = P(B)$.

Informally, two events are independent if they don't influence each other; knowing that A happened doesn't give you any additional information about B.

Independence Example - Discrete

- Suppose that our sample space consists of N flips of a coin that has probability p of giving heads.
- The events corresponding to a H in position i and in position j are independent.
- The chance of getting k heads in N flips is

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

The probability distribution on the set $0, \ldots, N$ given by this formula is called the *binomial distribution* for parameters p and N.

Independence Example - Continuous

Suppose we have a thermometer that measures the temperature t within an error $x = t - t_0$ from the true temperature, where x is normally distributed with standard deviation σ.

Suppose we make N independent measurements of the temperature. How are the errors distributed?

$$P(|x_1| < \delta, ..., |x_N| < \delta) =$$
$$\left(\frac{1}{\sigma\sqrt{2\pi}}\right)^N \int_{x_1=-\delta}^{\delta} ... \int_{x_N=-\delta}^{\delta} e^{-(\sum_{i=1}^{N} x_i^2)/(2\sigma^2)} dx_1 ... dx_N$$

This is the *multivariate gaussian* distribution.

Non-independent events

Suppose we draw a pair of real numbers (x, y) from the plane R^2 controlled by the distribution

$$P((x, y) \in U) = A \int_U e^{(-x^2 - xy - y^2)/(2\sigma^2)} dx dy$$

This density function has a bump at the origin and its level curves are ellipses.

The two coordinates are not independent of each other.

Non-independent events

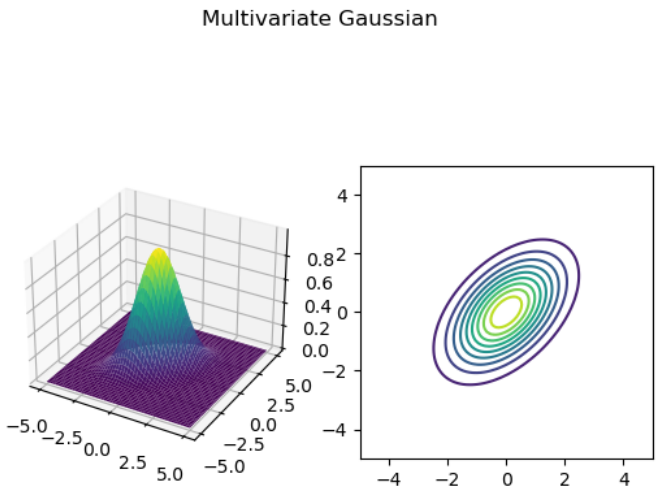


Figure 3: Multivariate Gaussian

Random Variables, Mean, and Variance

Random Variables

Definition: If X is a sample space, then a *random variable* is a real valued function $f : X \to \mathbf{R}$.

- Suppose that X is the sample space for a coin flip, so consists of heads and tails. Let $b(H) = 1$ and $b(T) = 0$. Then b is called a "Bernoulli Random Variable."
- For example, suppose that X is the set of N independent coin flips of a fair coin so that X consists of sequences of N heads or tails. If $x \in X$, let $f(x)$ be the number of heads. Then f is a random variable. Notice that f is the sum of N Bernoulli random variables.
- If X is a set of rolls of a pair of independent six-sided dice, and $f(x)$ is the sum of the values of the two dice, then f is another example of a random variable.
- If $U \subset \mathbf{R}$, and f is a random variable, then $f^{-1}(U) \subset X$ and, by definition,

$$P(f(x) \in U) = P(f^{-1}(U))$$

Example: For the coin-flipping example, suppose that $N = 4$ and f counts the number of heads. What is $P(f = 2)$?

Continuous random variable example

Suppose that $X = \mathbf{R}^2$ and

$$P(x \in U) = \left(\frac{1}{\sqrt{2\pi}}\right)^2 \int_U e^{-\|x\|^2/(2)} dxdy$$

Let $f(x) = \|x\|$. What is $P(f < r)$? In other words, how likely is a randomly drawn point to lie within distance r of the origin?

Independent random variables

Two random variables f and g are independent if:

- in the discrete case, $P(f = a \,\text{and}\, g = b) = P(f = a)P(g = b)$ for all $a, b \in \mathbf{R}$.
- in the continuous case, if $f^{-1}(U)$ and $g^{-1}(V)$ are independent for all intervals U and V in $\mathbf{R}$.

Expectation (Mean)

Definition: If f is a random variable on a sample space X, then

$$E[f] = \sum_{x \in X} f(x)P(x)$$

if X is discrete, or

$$E[f] = \int_X f(x)p(x)dx$$

where $p(x)$ is the density function, if X is continuous.

Properties:

- Linearity: $E[af + bg] = aE[f] + bE[g]$ if a and b are constants.
- If f and g are independent, then $E[fg] = E[f]E[g]$.

Example: If f is a binomial random variable with parameters N and p, then $E[f] = Np$.

Variance

Definition: If f is a random variable on a sample space X, then $\sigma^2(f)$, the *variance* of f is

$$\sigma^2(f) = E[(f - E[f])^2] = E[f^2] - E[f]^2$$

Variance of Binomial Random Variable

The variance of a binomial random variable with parameters N and p is $Np(1-p)$.

Variance of Normally Distributed Random Variable

The mean $E[x]$ of a normally distributed random variable with density

$$p(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-x^2/(2\sigma^2)}$$

is $E[x] = 0$.

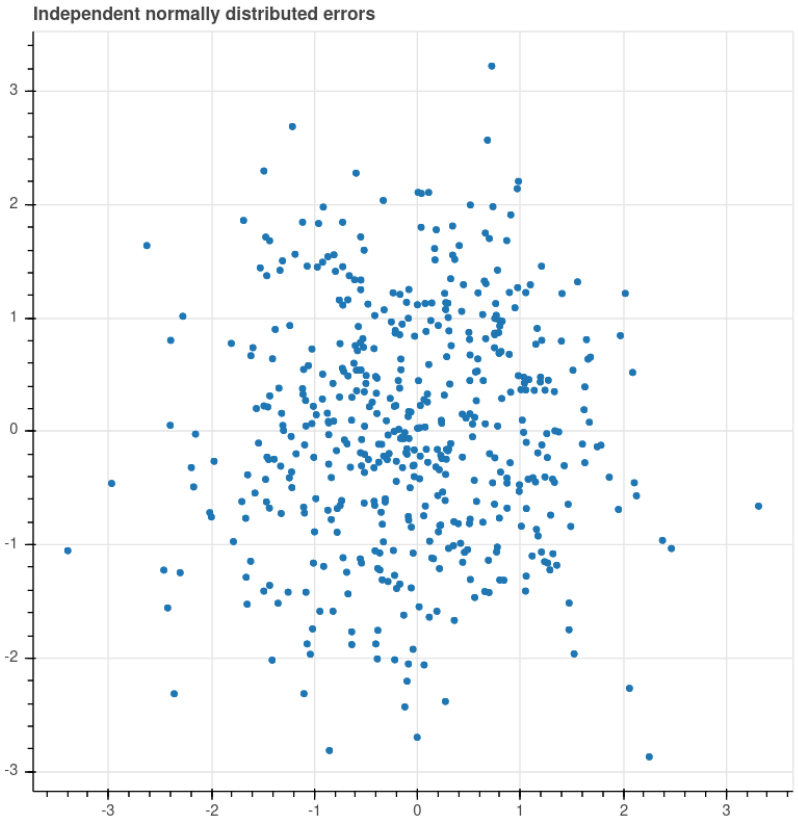
The variance $E[x^2] - E[x]^2 = E[x^2] = \sigma^2$.

Random Variables - continuous case

Random Variable: Continuous Case

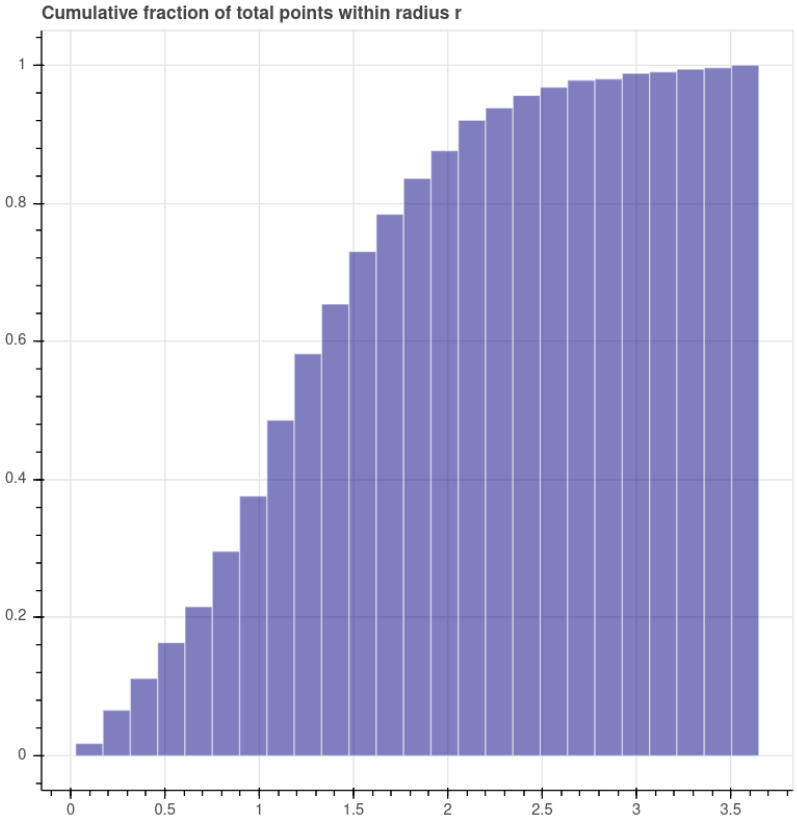
- We make two independent measurements of temperature t, where the true temperature is t_0 and the errors $x = t - t_0$ are independent normal random variables with variance 1.
- The sample space $X = \mathbf{R}^2$ and the probability density is the multivariate gaussian

$$p(x) = \frac{1}{2\pi} e^{(-x_1^2 - x_2^2)/2} = \frac{1}{2\pi} e^{-\|x\|^2/2}$$



Distribution of norms

- How is $\|x\| = \sqrt{x_1^2 + x_2^2}$ distributed? $\|x\|$ is a random variable on X.
- What is the probability $P(\|x\| < r)$?
- Here is a histogram using the sample data above showing the distribution of the distances. Notice that as r increases, more and more of the points lie within distance r of the origin.



Distribution of norms (continued)

- By definition,

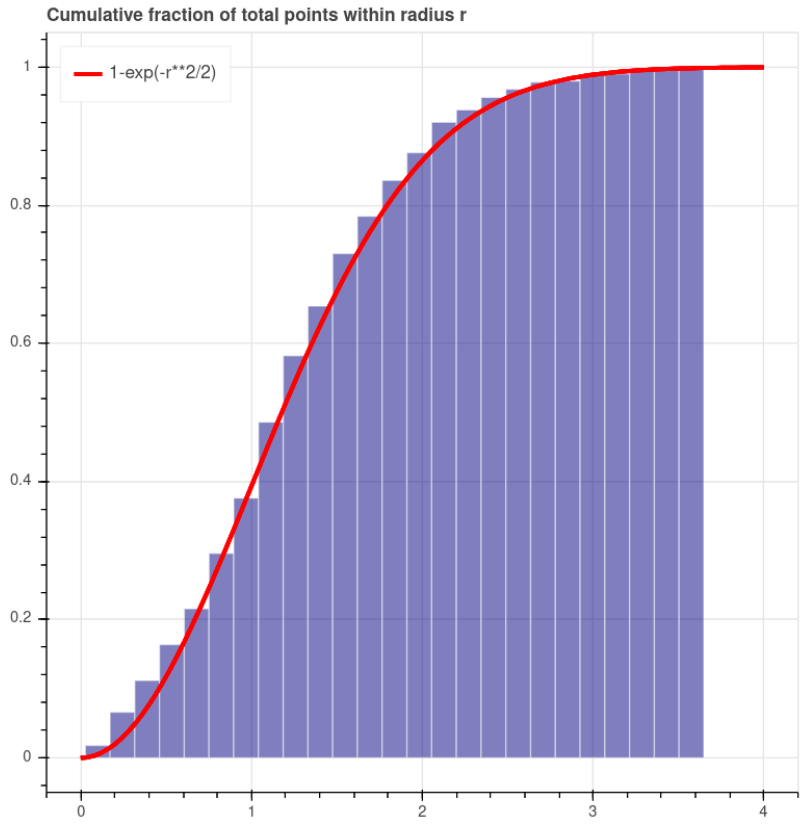
$$P(\|x\| < r) = P(\{(x_1, x_2) : x_1^2 + x_2^2 < r^2\}) = \frac{1}{2\pi} \int_{\|x\|<r} e^{-\|x\|^2/2}$$

- This is a doable integral using polar coordinates.

$$\begin{aligned} \frac{1}{2\pi} \int_{\|x\|<r} e^{-\|x\|^2/2} &= \frac{1}{2\pi} \int_{\theta=0}^{2\pi} \int_{\rho=0}^{r} e^{-\rho^2/2} \rho d\rho d\theta \\ &= \int_{\rho=0}^{r} \rho e^{-\rho^2/2} d\rho \\ &= 1 - e^{-r^2/2} \end{aligned}$$

Distribution of norms continued

- Here we superimpose our calculated cumulative density with the experimental data to see that they match.


Cumulative fraction of total points within radius r
1-exp(-r**2/2)
1
0.8
0.6
0.4
0.2
0
0
1
2
3
4

Models and Likelihood

Statistical Models

- Mathematical models
- Statistical models
  - Parameters
  - Likelihood

First example: coin flipping

- Model a coin flipping experiment as a Bernoulli random variable with parameter p.
- Flip the coin 100 times and get 55 heads and 45 tails.

$$L = \binom{100}{55} p^{55}(1-p)^{45}$$

- **Maximum Likelihood** - forget the constant as it doesn't effect the result.

$$\frac{dL}{dp} = 55p^{54}(1-p)^{45} - 45p^{55}(1-p)^{44} = 0$$

yields

$$55(1-p) = 45p$$

or

$$p = 55/100 = .55$$

Independent normally distributed errors

- Back to our temperature model. We assume that the errors in our measurements are normally distributed around zero. There is one parameter: the variance σ^2 in our density function for a single measurement

$$p_\sigma(x) = \left(\frac{1}{\sigma\sqrt{2\pi}}\right) e^{-x^2/(2\sigma^2)}$$

- We make N independent measurements of temperature

$$x_1, \dots, x_N$$

  What does this tell us about σ^2? The likelihood for independent measurements is the density

$$L = \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^N e^{-\|x\|^2/(2\sigma^2)}$$

- Maximize the density at this point. Use the *log-likelihood* as it is easier.

$$\log P(x) = -N \log \sigma - \frac{\|x\|^2}{2\sigma^2} + C$$

- Take the derivative and set it equal to zero.

- The *maximum likelihood estimate of the variance is the mean squared error!*

Linear Regression and likelihood

- Model says that our N data points (x_i, y_i) arose from a process

$$y = mx + b + \epsilon$$

  where ϵ is a normally distributed error term with variance σ^2.
- How should we set m, b, σ to make the observed data most likely?
- The density function is

$$P(m, b, \sigma) = \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^N e^{-\sum_{i=1}^{N}(y_i - mx_i - b)^2/(2\sigma^2)}$$

- The *log likelihood* is

$$\log P = -N \log \sigma - \frac{1}{2\sigma^2} \sum_i (y_i - mx_i - b)^2$$

- The derivatives with respect to m and b give the least squares estimates.

- The derivative with respect to σ gives the best estimate when σ^2 is the mean squared error.

- **Ordinary Least Squares is the maximum likelihood solution assuming independent normally distributed errors**

Bayesian Inference

Introduction

- Elements of Bayesian inference
  - a statistical model with parameters
  - a "prior distribution" on the parameters representing your state of knowledge about them
  - data arising from an experiment
  - an update to your prior distribution based on the experiment, leading to a "posterior distribution"

Example

- Rough example
  - you have a thermometer that reports the true temperature up to a normally distributed error. This is your statistical model.
  - you have a prior sense that the external temperature is around 30 degrees, based on the time of day and the time of year. This is your prior distribution.
  - you make several independent measurements using your thermometer, and it reports temperatures scattered around 40 degrees.
  - You conclude that the temperature is probably closer to 40 than 30 based on this data.

Bayes Theorem and Bayesian Inference

Suppose that t is the temperature and D is the data that is the result of our experiment. The heart of Bayesian inference is Bayes theorem:

$$P(t|D) = \frac{P(D|t)P(t)}{P(D)}$$

- $P(t|D)$ is the distribution of the temperature *given* the observed data.
- $P(D|t)$ is the probability that we would have observed the data, *given* what we know about the temperature.
- $P(t)$ is the *prior* distribution on the temperature.
- $P(D)$ is the probability of the data given all possible temperatures. Often it amounts to a constant that we can ignore.

More details in a specific case

We will use temperature measurements. There are two parameters: the true temperature t_* and the variance σ^2 of the errors in measurement. The probability density for our temperature measurements is the normal distribution

$$p(t) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(t-t_*)^2/(2\sigma^2)}$$

We don't know either the true temperature t_* or the variance σ^2.

We conduct an experiment and obtain temperature values $\mathbf{t}_0 = (t_1, \ldots, t_N)$.

Bayes Theorem in the temperature case

$$P(t_*, \sigma^2 | \mathbf{t} = \mathbf{t}_0) = \frac{P(\mathbf{t} = \mathbf{t}_0 | t_*, \sigma^2) P(t_*, \sigma^2)}{P(\mathbf{t} = \mathbf{t}_0)}$$

- The left hand side $P(t_*, \sigma^2 | \mathbf{t} = \mathbf{t}_0)$ is the *posterior distribution* and it is the distribution on t_* and σ^2 *given the results of our experiment.*
- The probability $P(\mathbf{t} = \mathbf{t}_0 | t_*, \sigma^2)$ is the *likelihood* that we would have obtained the data we got depending on the values of t_* and σ^2.
- The probability $P(t_*, \sigma^2)$ is the *prior distribution* that reflects our initial impression of the value of these parameters.
- The denominator $P(\mathbf{t} = \mathbf{t}_0)$ is the total probability of the results of the experiment:

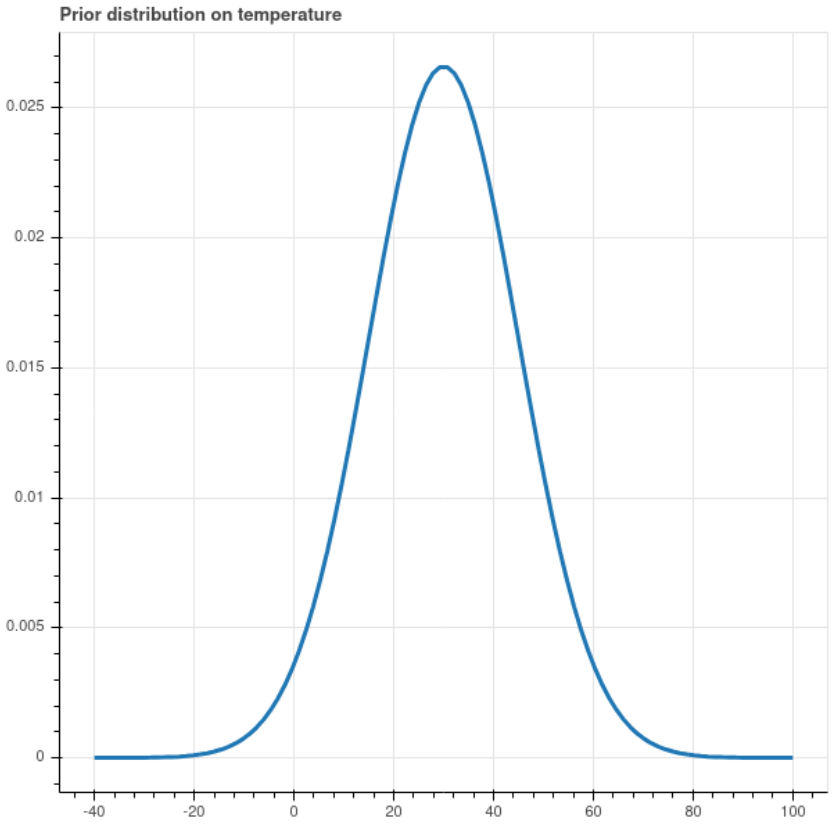
  $$P(\mathbf{t} = \mathbf{t}_0) = \int_{t_*, \sigma^2} P(\mathbf{t} = \mathbf{t}_0 | t_*, \sigma^2) P(t_*, \sigma^2)$$

  It functions as a normalizing constant and often we can get by without computing it at all.

Now we will do a worked example. We simplify the situation so that the variance in our thermometer $\sigma_*^2 = 1$.
Therefore the only unkown parameter is the true temperature t_*.

Prior Distribution

For our prior distribution, we assume that the average temperature is 30 degrees and the variance of is 15 degrees. That yields the following prior distribution.



The formula for this density is

$$P(t_*) = \frac{1}{\sqrt{30\pi}} e^{-(t_*-30)^2/30}$$

Likelihood of the data

We make independent measurements $\mathbf{t}_0 = (t_1, \dots, t_N)$.

The errors are $t_i - t_*$ where t_* is the true temperature. We have fixed the measurement variance at $\sigma^2 = 1$. Therefore

$$P(\mathbf{t} = \mathbf{t}_0 | t_*) = (\frac{1}{\sqrt{2\pi}})^N e^{-\|\mathbf{t}_0 - t_* \mathbf{e}\|^2/2}$$

where $\mathbf{e} = (1, 1, 1, \dots, 1)$.

The total probability

The total probability is the integral

$$P(\mathbf{t}_0) = \int_{t_*} P(\mathbf{t} = \mathbf{t}_0 | t_*) P(t_*)$$

Let's just call this T and avoid it for the moment.

Bayes Theorem

If we combine up all the constants in Bayes Theorem and call them A, we have

$$P(t_*|\mathbf{t} = \mathbf{t}_0) = Ae^{-\|\mathbf{t}-t_*\mathbf{e}\|^2/2-(t_*-30)^2/30}$$

The exponent in the exponential is

$$\begin{aligned} Q &= \|\mathbf{t} - t_*\mathbf{e}\|^2/2 + (t^* - 30)^2/30 \\ &= (t_* - 30)^2/30 + \sum_i (t_i - t_*)^2/2 \end{aligned}$$

By expanding this out and completing the square, you can show that

$$Q = (t_* - U)^2/2V + K$$

where K is a constant that doesn't involve t_*,

$$U = \frac{2 + \sum_i t_i}{\frac{1}{15} + N}$$

and

$$V = \frac{1}{\frac{1}{15} + N}$$

The posterior density

The previous calculation shows that the posterior density (up to multiplicative constants B) has the form

$$P(t_*|\mathbf{t} = \mathbf{t}_0) = Be^{-(t_*-U)^2/2V}$$

In other words, *it is a normal distribution centered at U with variance V.*

Suppose we measured temperatures

$$40, 41, 39, 37, 44.$$

Then $N = 5$, the mean of these observations is 40.2 and the variance is 5.4

We have

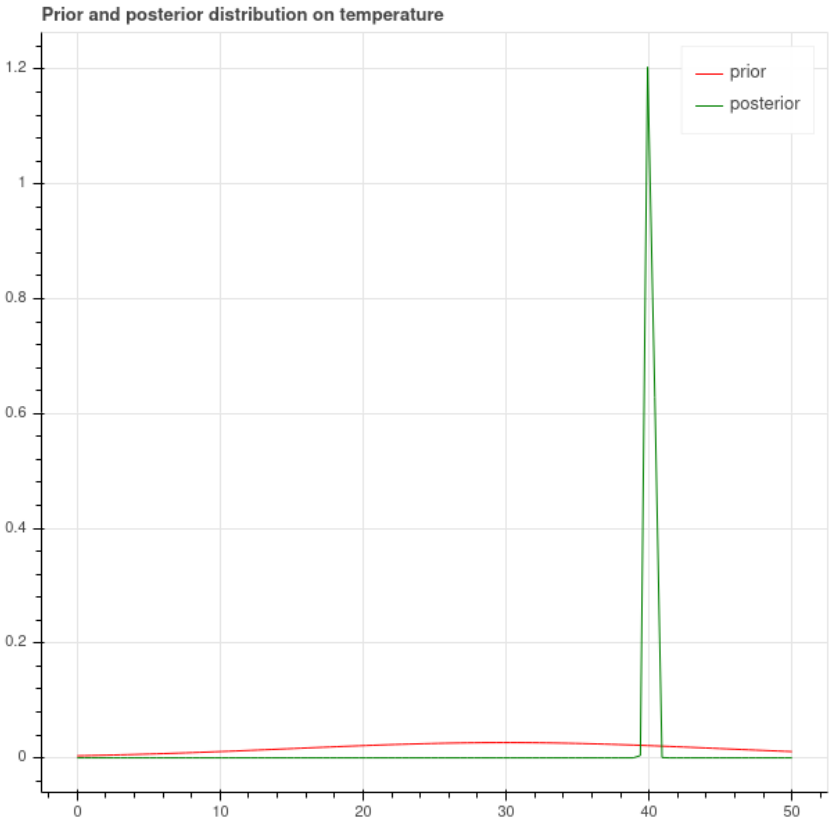
$$U = 40.1$$

and

$$V = 0.2$$

. The posterior mean is a bit less than the observed mean because our prior pulls it towards 30.

Notice in the formula for U and V that as $N \to \infty$ the posterior mean U approaches the sample mean $\frac{1}{N}\sum t_i$ and the variance approaches 0.



General Result

Proposition: Suppose that our statistical model for an experiment proposes that the measurements are normally distributed around an unknown mean value of μ with a fixed, known variance of σ^2. Suppose that our prior distribution on μ is also normal with mean μ_0 and variance τ^2. Finally imagine that we make measurements

$$y_1, \cdots, y_N.$$

The posterior distribution on μ is again normal, with posterior variance

$$\tau'^2 = \frac{1}{\frac{1}{\tau^2} + \frac{N}{\sigma^2}}$$

and posterior mean

$$\mu' = \frac{\frac{\mu_0}{\tau^2} + \frac{N}{\sigma^2}\overline{y}}{\tau'^2}$$

So the posterior mean is a weighted average of the sample mean and the prior mean, and as $N \to \infty$, the posterior mean approaches the sample mean and the prior has less and less influence on the interpretation of the experiment.

Bayesian Coin Flipping

Elements of Bayesian inference

We return to the coin flipping experiment. The ingredients of our Bayesian analysis of this situation are:

- a statistical model. We assume that our coin is modelled by a Bernoulli random variable with parameter p of returning heads. The likelihood of getting h heads in N flips is given by the binomial distribution

$$P(h|p) = \binom{N}{h} p^h (1-p)^{N-h}.$$

- a prior distribution $P(p)$. Initially, we make no assumptions about the coin, so we choose the *uniform distribution* that assigns probability density 1 to every $p \in [0, 1]$.
- some data D. We flip the coin N times and receive h heads; that's our data.

Our problem is to construct a posterior distribution $P(p|h)$ that tells us how this experiment updates our impressions about the coin.

Bayes's theorem

From our setup and Bayes's theorem:

$$P(p|h) = \frac{P(h|p)P(p)}{P(h)} = \frac{\binom{N}{h}p^h(1-p)^{N-h}}{P(h)}$$

where the denominator is

$$P(h) = \binom{N}{h}\int_{p=0}^{1} p^h(1-p)^{N-h}dp$$

The posterior

The posterior distribution, up to a constant A, is

$$P(p|h) = Ap^h(1-p)^{N-h}$$

We know from our discussion of maximum likelihood that the *most likely* value of p is h/N, the fraction of heads among all flips. This is called the *maximum a posteriori estimate* or MAP.

The posterior mean

In Bayesian inference, one often uses the *mean of the posterior distribution* as a better summary of the posterior than the point where the posterior is a maximum. To compute the mean, we need to know the constant A, which is

$$A = \frac{1}{\int_{p=0}^{1} p^h (1-p)^{N-h} dp}$$

The mean of the posterior is given by the formula

$$E[p|h] = A \int_{p=0}^{1} p^{h+1} (1-p)^{N-h} dp$$

The *Beta Integral* is the integral

$$B(a,b) = \int_{p=0}^{1} p^{a-1} (1-p)^{b-1} dp$$

and with some work one can show that

$$B(a,b) = \frac{a+b}{ab} \frac{1}{\binom{a+b}{a}}$$

Putting this all together gives the result

$$E[p|h] = \frac{h+1}{N+2}$$

Some numbers

- Given 55 heads out of 100 flips, the maximum likelihood estimate for p (and the maximum a posteriori estimate assuming a uniform prior) is $p = .55$.
- The posterior mean is $56/102 = .549$ which is a bit less.