THE UNIVERSITY OF TEXAS AT AUSTIN
Department of Statistics and Data Sciences
College of Natural Sciences

SDS 321: Introduction to Probability and Statistics

Probability and Statistics-review

Purnamrita Sarkar
Department of Statistics and Data Sciences
The University of Texas at Austin

www.cs.cmu.edu/ psarkar/teaching

Things you should know

- **Single r.v.**
  - How to get a marginal pdf from a joint.
  - How to calculate expectation, variance of a r.v.
  - Derived distributions. Pay special attention to the Uniform. Its always tricky,
- **Multiple r.v.'s**
  - When are two r.v.'s independent?
  - How to calculate covariance, correlation of two r.v's
  - Conditional expectation, iterated expectation rule.
  - Linearity of expectation. Variance of sum of r.v.'s
  - Covariance of two sums of r.v's
- Markov's inequality, Chebyshev's inequality. Central limit theorem.

Functions of a continuous random variable: Summary

- We know that functions $Y = g(X)$ of random variables X are again random variables.
- In the discrete case, it was pretty easy to find the PMF of the new random variable X:

$$p_Y(y) = \sum_{x|g(x)=y} p_X(x)$$

- In the continuous case, we need to find $F_Y(y) = \mathbf{P}(Y \leq y) = \int_{x|g(X)\leq y} f_X(x)dx$
- We can then differentiate wrt y to get the PDF of Y.
- In certain cases, this procedure simplifies a little:
- **Linear case**: If $Y = aX + b$, $f_Y(y) = \frac{1}{|a|} f_X\left(\frac{y-b}{a}\right)$
- **Monotonic case**: If g is a strictly monotonic function with inverse h, then $f_Y(y) = f_X(h(y))\left|\frac{dh}{dy}(y)\right|$
- What about if we have functions of more than one random variable?

Functions of continuous random variables: Example

- What is the PDF of the area of a circle, if the radius is a uniform random variable on [0,1]?
- If X is the radius, we know that the area $Y = \pi X^2$.
- We know that

$$f_X(x) = \begin{cases} 1 & 0 \leq x \leq 1 \\ 0 & \text{otherwise} \end{cases}$$

Functions of continuous random variables: Example

- What is the PDF of the area of a circle, if the radius is a uniform random variable on [0,1]?
- If X is the radius, we know that the area $Y = \pi X^2$.
- We know that

$$f_X(x) = \begin{cases} 1 & 0 \leq x \leq 1 \\ 0 & \text{otherwise} \end{cases}$$

- We can calculate the CDF of Y:

$$\begin{aligned} F_Y(y) =& \mathbf{P}(\pi X^2 \leq y) \\ =& \int_{x|\pi x^2 \leq y} f_X(x)dx \end{aligned}$$

Functions of continuous random variables: Example

- What is the PDF of the area of a circle, if the radius is a uniform random variable on [0,1]?
- If X is the radius, we know that the area $Y = \pi X^2$.
- We know that

$$f_X(x) = \begin{cases} 1 & 0 \leq x \leq 1 \\ 0 & \text{otherwise} \end{cases}$$

- We can calculate the CDF of Y:

$$\begin{aligned} F_Y(y) =& \mathbf{P}(\pi X^2 \leq y) \\ =& \int_{x|\pi x^2 \leq y} f_X(x) dx \end{aligned}$$

- $f_X(x)$ is only positive for $0 \leq x \leq 1$, corresponding to $0 \leq y \leq \pi$. So,

$$F_Y(y) = \begin{cases} 0 & y < 0 \\ \int_0^{\sqrt{y/\pi}} f_X(x) dx = \sqrt{\frac{y}{\pi}} & 0 \leq y \leq \pi \\ 1 & y > \pi \end{cases}$$

Functions of continuous random variables: Example

- From the previous slide:

$$F_Y(y) = \begin{cases} 0 & y < 0 \\ \int_0^{\sqrt{y/\pi}} f_X(x)dx = \sqrt{\frac{y}{\pi}} & 0 \leq y \leq \pi \\ 1 & y > \pi \end{cases}$$

- We can differentiate to get

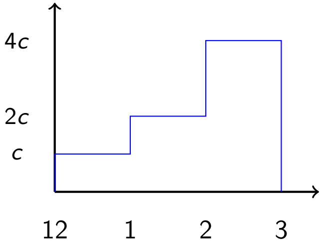
$$f_Y(y) = \begin{cases} 0 & y < 0 \\ \frac{d}{dy}\left(\sqrt{\frac{y}{\pi}}\right) = \frac{1}{2}\sqrt{\frac{1}{y\pi}} & 0 \leq y \leq \pi \\ 0 & y > \pi \end{cases}$$

Conditional expectation

- I am expecting an email, that will definitely arrive between midday and 3pm.
- Within a given hour (midday-1, 1-2, 2-3), each time is equally likely.
- It is twice as likely to arrive between 1 and 2 as it is to arrive between midday and 1.
- It is twice as likely to arrive between 2 and 3 as it is to arrive between 1 and 2.
- What does the PDF look like?

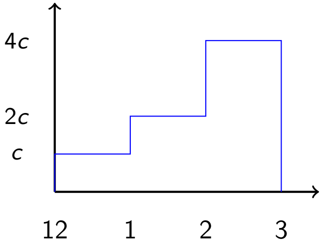
Conditional expectation

- I am expecting an email, that will definitely arrive between midday and 3pm.
- Within a given hour (midday-1, 1-2, 2-3), each time is equally likely.
- It is twice as likely to arrive between 1 and 2 as it is to arrive between midday and 1.
- It is twice as likely to arrive between 2 and 3 as it is to arrive between 1 and 2.
- What does the PDF look like?



Conditional expectation

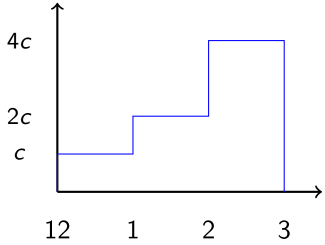
- I am expecting an email, that will definitely arrive between midday and 3pm.
- Within a given hour (midday-1, 1-2, 2-3), each time is equally likely.
- It is twice as likely to arrive between 1 and 2 as it is to arrive between midday and 1.
- It is twice as likely to arrive between 2 and 3 as it is to arrive between 1 and 2.
- What does the PDF look like?



- What is c?

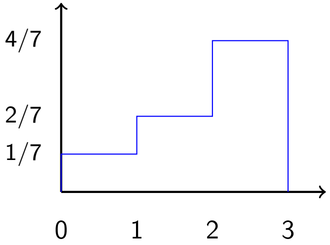
Conditional expectation

- I am expecting an email, that will definitely arrive between midday and 3pm.
- Within a given hour (midday-1, 1-2, 2-3), each time is equally likely.
- It is twice as likely to arrive between 1 and 2 as it is to arrive between midday and 1.
- It is twice as likely to arrive between 2 and 3 as it is to arrive between 1 and 2.
- What does the PDF look like?



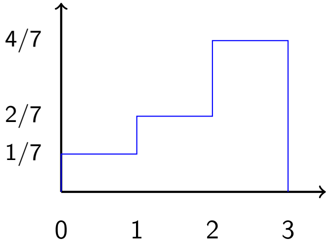
- What is c? $1/7$

Conditional expectation



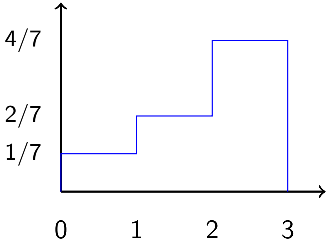
- I wait until 2pm. It still hasn't arrived. What is the expected value of the arrival time?
- What is the expected time without any conditioning?

Conditional expectation



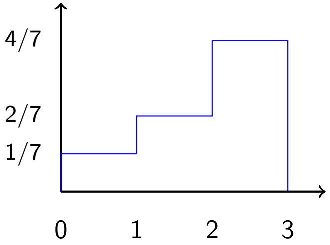
- I wait until 2pm. It still hasn't arrived. What is the expected value of the arrival time?
- What is the expected time without any conditioning?

Conditional expectation



- I wait until 2pm. It still hasn't arrived. What is the expected value of the arrival time?
- What is the expected time without any conditioning?
- First, what is the conditional probability, $f_{X|X>2}(x)$?

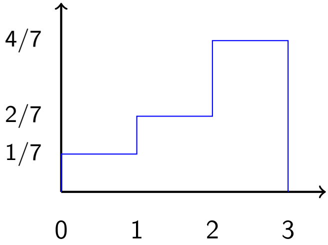
Conditional expectation



- I wait until 2pm. It still hasn't arrived. What is the expected value of the arrival time?
- What is the expected time without any conditioning?
- First, what is the conditional probability, $f_{X|X>2}(x)$?

$$f_{X|X>2}(x) = \begin{cases} 1 & \text{if } 2 < x \leq 3 \\ 0 & \text{otherwise} \end{cases}$$

Conditional expectation

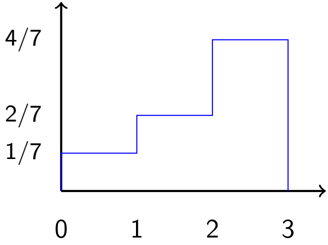


- I wait until 2pm. It still hasn't arrived. What is the expected value of the arrival time?
- What is the expected time without any conditioning?
- First, what is the conditional probability, $f_{X|X>2}(x)$?

$$f_{X|X>2}(x) = \begin{cases} 1 & \text{if } 2 < x \leq 3 \\ 0 & \text{otherwise} \end{cases}$$

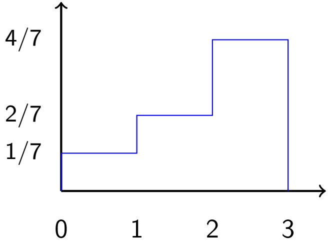
- So, $E[X|X > 2] = \int_{-\infty}^{\infty} x f_{X|X>2}(x) dx = \int_2^3 x\, dx = 2.5.$

Conditional expectation



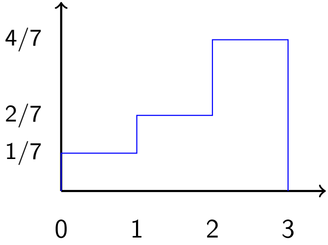
- What is the (unconditional) probability that $X > 2$?

Conditional expectation



- What is the (unconditional) probability that $X > 2$?
- $\mathbf{P}(X > 2) = \int_2^3 f_X(x)dx = 4/7$

Conditional expectation



- What is the (unconditional) probability that $X > 2$?
- $\mathbf{P}(X > 2) = \int_2^3 f_X(x)dx = 4/7$
- Similarly, $\mathbf{P}(X < 1) = \int_0^1 f_X(x)dx = 1/7$ and $\mathbf{P}(1 \leq X \leq 2) = 2/7$.

Total expectation theorem

- What is the total expectation of X?

Total expectation theorem

- What is the total expectation of X?
- $E[X] = \int_{-\infty}^{\infty} x f_X(x) dx$

Total expectation theorem

- What is the total expectation of X?
- $E[X] = \int_{-\infty}^{\infty} x f_X(x) dx$

Total expectation theorem

- What is the total expectation of X?
- $E[X] = \int_{-\infty}^{\infty} x f_X(x) dx$
- Total expectation theorem gives:

$$\begin{aligned} E[X] =& E[X|X \leq 1]\mathbf{P}(X \leq 1) + E[X|1 \leq X \leq 2]\mathbf{P}(1 \leq X \leq 2) \\ &+ E[X|X > 2]\mathbf{P}(X > 2) \\ =& 0.5 \cdot 1/7 + 1.5 \cdot 2/7 + 2.5 \cdot 4/7 = 27/14 \end{aligned}$$

Independent r.v's

A man and a woman agree to meet at a certain location at about 12:30 p.m. Suppose the man arrives at a time uniformly distributed between 12:00 and 12:45, and the woman independently arrives at a time uniformly distributed between 12:15 and 1:00. Let X be the man's arrival time and Y be the woman's arrival time.

- Find the probability that the first to arrive shows up before $12:20$.
- Find the probability that the first to arrive waits no longer than 5 minutes.
- What is the probability that the man arrives first?

Independent r.v's

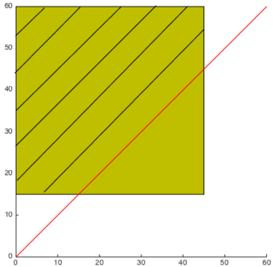
Find the probability that the first to arrive shows up before 12 : 20.

$$
\begin{aligned}
P(\min(X,Y) \le 12:20) &= 1 - P(\min(X,Y) \ge 12:20) \\
&= 1 - P(X \ge 12:20)P(Y \ge 12:20) \\
&= 1 - (1 - P(X \le 12:20))(1 - P(Y \le 12:20)) \\
&= 1 - (1 - \frac{5}{45})(1 - \frac{20}{45}) \\
&= \frac{41}{81}
\end{aligned}
$$

Independent r.v's

Find the probability that the man arrives first.

$$P(X < Y) = 1 - 1/45^2 \times .5 \times 30^2 = 1 - \frac{2}{9} = 7/9$$



Derived distributions

- We know that a function of a random variable is a random variable.
- We know that a function of two random variables is a random variable.
- Sometimes, we need to figure out the distribution of that random variable!
- This is often pretty easy in the discrete case.
- e.g. I roll two fair dice. Let X be the number on the first one, and Y be the number on the second one.
- Let $Z = XY$. What is the PMF of Z?
- We can just enumerate through all the options!

Derived distributions

- $\mathbf{P}(Z = 1) = \mathbf{P}(X = 1)\mathbf{P}(Y = 1) = 1/36$
- $\mathbf{P}(Z = 2) = \mathbf{P}(X = 1)\mathbf{P}(Y = 2) + \mathbf{P}(X = 2)\mathbf{P}(Y = 1) = 2/36$
- $\mathbf{P}(Z = 3) = \mathbf{P}(X = 1)\mathbf{P}(Y = 3) + \mathbf{P}(X = 3)\mathbf{P}(Y = 1) = 2/36$
- $\mathbf{P}(Z = 4) = \mathbf{P}(X = 1)\mathbf{P}(Y = 4) + \mathbf{P}(X = 2)\mathbf{P}(Y = 2) + \mathbf{P}(X = 4)\mathbf{P}(Y = 1) = 3/36 \ldots$

- For general $Z = z$, we can write:

$$\mathbf{P}(Z = z) = \sum_{x=1}^{6} \mathbf{P}(X = x)\mathbf{P}(Y = z/x)$$

Independence, joint and marginal

- $f_{X,Y}(x,y) = ce^{-(x+y^2)/2}$ for $x \geq 0 - \infty < y < \infty$
- What is c?
- Are X, Y independent?
- What is $\text{var}(X+Y)$?
- What is $\text{cov}(X,Y)$

Independence, joint and marginal

- Seems like its a product of an exponential and a normal with 0 mean variance 1.
- $f_X(x) = .5e^{-.5x}$
- $f_Y(y) = 1/\sqrt{2\pi}e^{-y^2/2}$
- $c = \frac{.5}{\sqrt{2\pi}}$
- Are X, Y independent? Yes!
- What is $\text{var}(X+Y) = \text{var}(X) + \text{var}(Y) = 2^2 + 1 = 5$.
- What is $\text{cov}(X, Y)$? 0

Conditional expectation: Conditioning on a concrete event

- A bird leaves its nest heading north, at 1pm. Let X be the speed at which a bird flies, and Y be the bird's distance north from the nest at 1:30pm.
- Assume $X \sim \mathit{Uniform}(3, 6)$ and $Y|X = x \sim \mathit{Normal}(x/2, 1)$
- What is $E[Y]$? What is $\mathrm{var}(Y)$?
- $E[Y|X = x] = x/2$ and so $E[Y|X] = X/2$ and so $E[Y] = E[E[Y|X]] = E[X/2] = 4.5$

Covariance and Correlation

Assume that X and Y are correlated with $E[X] = 1$, $E[Y] = 2$, $E[X^2] = 5$.

- $\text{var}(X) =$
- $\text{var}(Y) =$
- $\text{cov}(X, Y) =$
- $\rho(X, Y) =$

Covariance

Assume that X and Y are correlated with $E[X] = 1$, $E[Y] = 2$, $E[X^2] = 5$, $E[Y^2] = 20$,and $E[XY] = 3$

- $\text{var}(X) = 5 - 1 = 4$
- $\text{var}(Y) = 20 - 4 = 16$
- $\text{cov}(X, Y) = E[XY] - E[X]E[Y] = 3 - 2 = 1$
- $\rho(X, Y) = 1/\sqrt{4 \times 16} = 1/8$

Assume that X and Y are correlated with $E[X] = 1$, $E[Y] = 2$, $E[X^2] = 5$.

- $\text{var}(X) =$
- $\text{var}(Y) =$
- $\text{cov}(X, Y) =$
- $\rho(X, Y) =$

Statistics

- Inequalities: Markov and Chebyshev
- We have used the Weak Law of Large numbers and the Central Limit theorem.
- In Frequentist statistics, we've looked at **the MLE**, **Type I error** and **significance testing**.
- Today, we will review these with some examples.

Frequentist statistics

- The frequentist viewpoint is that probability $=$ frequency.
- Frequentists treat parameters as fixed (they don't change if we repeat the experiment) and observations as random (they do change if we repeat the experiment).
- Frequentists use **estimators** to estimate the parameters, based on some sequence of n observations.
- These estimators might be **unbiased** (right on average for any number of observations), **asymptotically unbiased** (right on average as $n \to \infty$), or **biased** (wrong on average).
- In general, if we decrease the bias of an estimator, we increase it's variance.

The Maximum Likelihood estimator

You have n i.i.d. datapoints $X_1, \dots X_n \sim N(\mu, \sigma_1^2)$ and n i.i.d. datapoints $Y_1, \dots Y_n \sim N(\mu, \sigma_2^2)$. Whats the MLE of μ? You can assume that you know σ_1^2 and σ_2^2.

- Write the log likelihood: $f(x_1, \dots, x_n, y_1, \dots, y_n) = \prod_i \exp(-(x_i - \mu)^2/2\sigma_1^2) \prod_j \exp(-(y_j - \mu)^2/2\sigma_2^2)$
- Take log: $L = -\sum_i \frac{(x_i - \mu)^2}{2\sigma_1^2} - \sum_j \frac{(y_j - \mu)^2}{2\sigma_2^2}$
- Differentiate: $\frac{dL}{d\mu} = -\sum_i \frac{x_i - \mu}{\sigma_1^2} - \sum_j \frac{y_j - \mu}{\sigma_2^2}$
- Set to zero and solve for $\hat{\mu}$. $\hat{\mu} = \frac{\bar{x}/\sigma_1^2 + \bar{y}/\sigma_2^2}{1/\sigma_1^2 + 1/\sigma_2^2}$

The Maximum Likelihood estimator

- Is this biased?
- $E[\hat{\mu}] = \dfrac{\mu/\sigma_1^2 + \mu/\sigma_2^2}{1/\sigma_1^2 + 1/\sigma_2^2} = \mu$

Confidence intervals

- Usually, our estimators will correspond (exactly or approximately – central limit theorem!) to the sample mean of a normal distribution.
- If our estimator is the sample mean of n samples from a normal distribution with known variance σ^2, the sample mean is distributed according to
$$\bar{X} \sim Normal\left(\mu, \frac{\sigma^2}{n}\right)$$
- So, the difference between the true mean and the sample mean can be described as
$$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim Normal(0, 1)$$
- The $(1 - \alpha)$ confidence region for the true mean corresponds to the region where
$$\mathbf{P}\left(\frac{|\bar{X} - \mu|}{\sigma/\sqrt{n}} \leq z\right) = 1 - \alpha$$
- We can find an appropriate value for z from our standard normal tables.
- Then, our $(1 - \alpha)$ confidence region is $[\bar{X} - z\sigma/\sqrt{n}, \bar{X} + z\sigma/\sqrt{n}]$

Confidence intervals: unknown variance

- If our estimator is the sample mean of n samples from a normal distribution with unknown mean and variance, the distribution over $\bar{X}$ is no longer normal.
- However we still know it's distribution!
- If we knew σ, then $\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}$ would be a standard normal random variable.
- Let $\hat{s}_{n-1}^2 = \sum_i (x_i - \bar{X})^2/(n-1)$.
- Then $T_n = \frac{\bar{X}-\mu}{\hat{s}_{n-1}/\sqrt{n}}$ follows a t-distribution with $(n-1)$ degrees of freedom.
- We can look up the corresponding probabilities using the t-distribution table.

Exponential confidence interval

I have $X_1, \ldots, X_n \sim Exp(\lambda)$. Find a 95% confidence interval for λ

- Start with what we know.
- We want θ^- and θ^+ such that $P(\theta^- \leq \lambda \leq \theta^+) = .95$
- But $\bar{X} \sim N(1/\lambda, 1/(n\lambda^2))$ as n gets large by CLT
- $P\left(\sqrt{n}\frac{|\bar{X} - 1/\lambda|}{1/\lambda} \leq 1.96\right) = 0.95$
- What does this translate to?
- $P\left(\frac{1}{\lambda}\left(1 - \frac{1.96}{\sqrt{n}}\right) \leq \bar{X} \leq \frac{1}{\lambda}\left(1 + \frac{1.96}{\sqrt{n}}\right)\right) = .95$
- $P\left(\frac{1}{\bar{X}}\left(1 - \frac{1.96}{\sqrt{n}}\right) \leq \lambda \leq \frac{1}{\bar{X}}\left(1 + \frac{1.96}{\sqrt{n}}\right)\right) = .95$

Hypothesis testing

I have two datasets, $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_n$, with known variance 1 and unknown means μ_X and μ_Y. I want to test the null hypothesis that $\mu_X = 2\mu_Y$ and $H_1 : \mu_X \neq 2\mu_Y$ using a two-sided significance test.

- Suggest an appropriate statistic – specifically, a function of $\bar{X}$ and $\bar{Y}$ – that has zero mean under the null hypothesis. $S = \bar{X} - 2\bar{Y}$?
- What is the variance of this statistic? $\text{var}(S) = 5/n$
- Construct an appropriate significance test for the null hypothesis, at significance level $\alpha = 0.05$
  $P(|S| \geq \xi) = P(|S|\sqrt{n}/\sqrt{5} \geq \xi\sqrt{n}/\sqrt{5}) = .95$ So $\xi = 1.96\sqrt{5/n}$

What you should know

- **Frequentist Statistics**
  - Markov, Chebyshev inequalities and when to use them?
  - Weak Law of Large numbers, and Central limit theorem.
  - What are biased/unbiased/asymptotically unbiased estimators?
  - How to get the MLE. Pay special attention to the Uniform.
  - How to get a confidence interval with a given coverage probability? Normal vs t tables.
  - For any rejection region how to get type I error.
  - How to get the rejection region with a fixed type I error?
  - How to do a significance test with a given level of significance?