W1211 Introduction to Statistics
Lecture 10

Wei Wang

Oct 8, 2012

Random Variables v.s. Distributions

- Distributions are property of Random Variables, which gives provides probabilistic description of RVs.
- An RV only has one distribution.
- Two RVs can have the same distribution.

Poisson Distribution

- Poisson Distribution is for describing outcomes that come in the form of count data, e.g., visits to a particular website during a time interval
- But unlike Binomial or Hypergeometric Distribution, there is no simple experiment that Poisson Distribution is based on.
- A random variable X is said to have Poisson Distribution with parameter $\mu (> 0)$ if the pmf of X is

$$p(x;\mu) = e^{-\mu}\frac{\mu^x}{x!}, x = 0, 1, 2, \ldots$$

Poisson Distribution PMF

- Verify the pmf is a valid pmf

$$p(x; \mu) = e^{-\mu} \frac{\mu^x}{x!}, x = 0, 1, 2, \ldots$$

Poisson Distribution PMF

- Verify the pmf is a valid pmf

$$p(x;\mu) = e^{-\mu}\frac{\mu^x}{x!}, x = 0, 1, 2, \ldots$$

- Recall from Calculus

$$e^{\mu} = 1 + \mu + \frac{\mu^2}{2!} + \frac{\mu^3}{3!} + \frac{\mu^4}{4!} + \cdots$$

Poisson Distribution PMF

- Verify the pmf is a valid pmf

$$p(x;\mu) = e^{-\mu}\frac{\mu^x}{x!}, x = 0, 1, 2, \ldots$$

- Recall from Calculus

$$e^{\mu} = 1 + \mu + \frac{\mu^2}{2!} + \frac{\mu^3}{3!} + \frac{\mu^4}{4!} + \cdots$$

- So

$$p(0;\mu) + p(1;\mu) + p(2;\mu) + \cdots = e^{\mu} \times e^{-\mu} = 1$$

Example

- Let X denote the number of creatures of a particular type captured in a trap during a given time period. Suppose that X has a Poisson distribution with =4.5, so on average traps will contain 4.5 creatures. Then the probability that a trap contains exactly five creatures is

$$P(X = 5) = \frac{e^{-4.5}(4.5)^5}{5!} = 0.1708$$

Example

- Let X denote the number of creatures of a particular type captured in a trap during a given time period. Suppose that X has a Poisson distribution with =4.5, so on average traps will contain 4.5 creatures. Then the probability that a trap contains exactly five creatures is

$$P(X = 5) = \frac{e^{-4.5}(4.5)^5}{5!} = 0.1708$$

- The probability that the a trap has at most five creatures is

$$P(X \leq 5) = \sum_{x=0}^{5} \frac{e^{-4.5}(4.5)^x}{x!} = .7029$$

Poisson Distribution as a Limit

- Suppose that in the binomial pmf $b(x; n; p)$, we let $n \to \infty$ and $p \to 0$ in such a way that np approaches a value $\mu > 0$. Then $b(x; n; p) \to p(x; \mu)$.
- So in any binomial experiment in which n is large and p is small, , then Binomial can be approximated by Poisson Distribution with parameter $\mu = np$.

Example

- A typesetter, on the average makes one error in every 500 words typeset. A typical page contains 300 words. What is the probability that there will be no more than two errors in five pages?

Example

- A typesetter, on the average makes one error in every 500 words typeset. A typical page contains 300 words. What is the probability that there will be no more than two errors in five pages?
- X is the number of errors in five pages

$$X \sim Bin(1500, 1/500)$$

Example

- A typesetter, on the average makes one error in every 500 words typeset. A typical page contains 300 words. What is the probability that there will be no more than two errors in five pages?
- X is the number of errors in five pages

$$X \sim Bin(1500, 1/500)$$

- Exact solution

$$P(X \leq 2) = \sum_{x=0}^{2} \binom{1500}{x} (\frac{1}{500})^{x} (\frac{499}{500})^{1500-x} = .4230$$

Example

- A typesetter, on the average makes one error in every 500 words typeset. A typical page contains 300 words. What is the probability that there will be no more than two errors in five pages?
- X is the number of errors in five pages

$$X \sim Bin(1500, 1/500)$$

- Exact solution

$$P(X \leq 2) = \sum_{x=0}^{2} \binom{1500}{x} (\frac{1}{500})^x (\frac{499}{500})^{1500-x} = .4230$$

- With Poisson Approximation $\mu = np = 3$

$$P(X \leq 2) \approx e^{-3} + 3e^{-3} + \frac{3^2 e^{-3}}{2} = .4232$$

Mean and Variance of Poisson Distribution

- If X has a Poisson Distribution with parameter μ, then $E(X) = Var(X) = \mu$.
- It can be derived directly from the pmf, or through the Binomial limit argument.
- If X is $b(x; n; p)$, then

$$E(X) = np \to \mu, \; Var(X) = np(1-p) \to \mu$$

The Poisson Process

- The Poisson RV(Random Variable) describes the count in a fixed time period. We can further consider how the RV changes over time.

The Poisson Process

- The Poisson RV(Random Variable) describes the count in a fixed time period. We can further consider how the RV changes over time.
- Suppose we are observing occurrence of a type of events, let $P_k(t)$ denote the probability that k events will be observed during any particular time interval of length t, then if

$$P_k(t) = e^{-\alpha t} \frac{(\alpha t)^k}{k!}$$

the we say the events occur according to a Poisson Process with rate α.

Continuous Random Variables

Continuous RV

- Recall the definition of pmf for a discrete rv. $P(X=x)$. Can we extend this definition to continuous rv's?

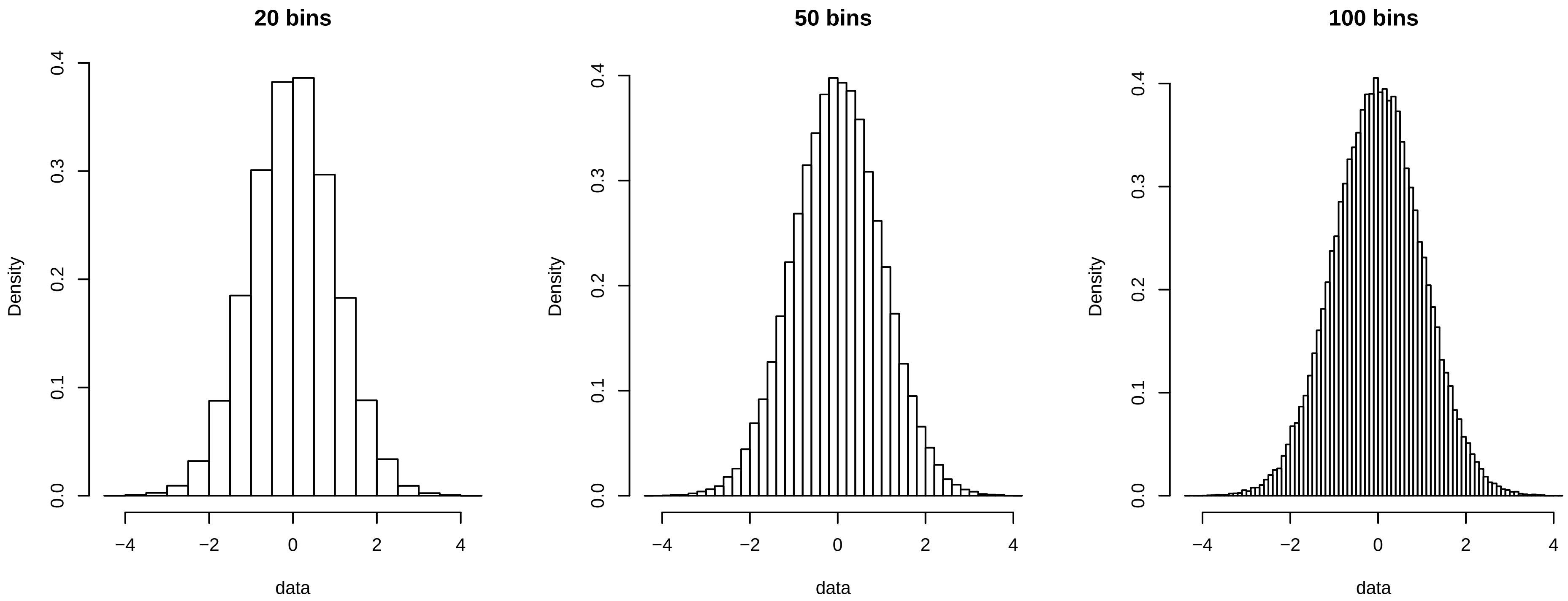
- Uniform random variable: X is equally likely to be any number on [0,1], what is the probability $P(X=0.5)$?

- The probability model for a continuous random variable assigns probabilities to intervals of outcomes rather than to individual outcomes.

- The probability model of X is often described by a smooth curve, which is the probability density function (pdf) of X.

From Histogram to Density

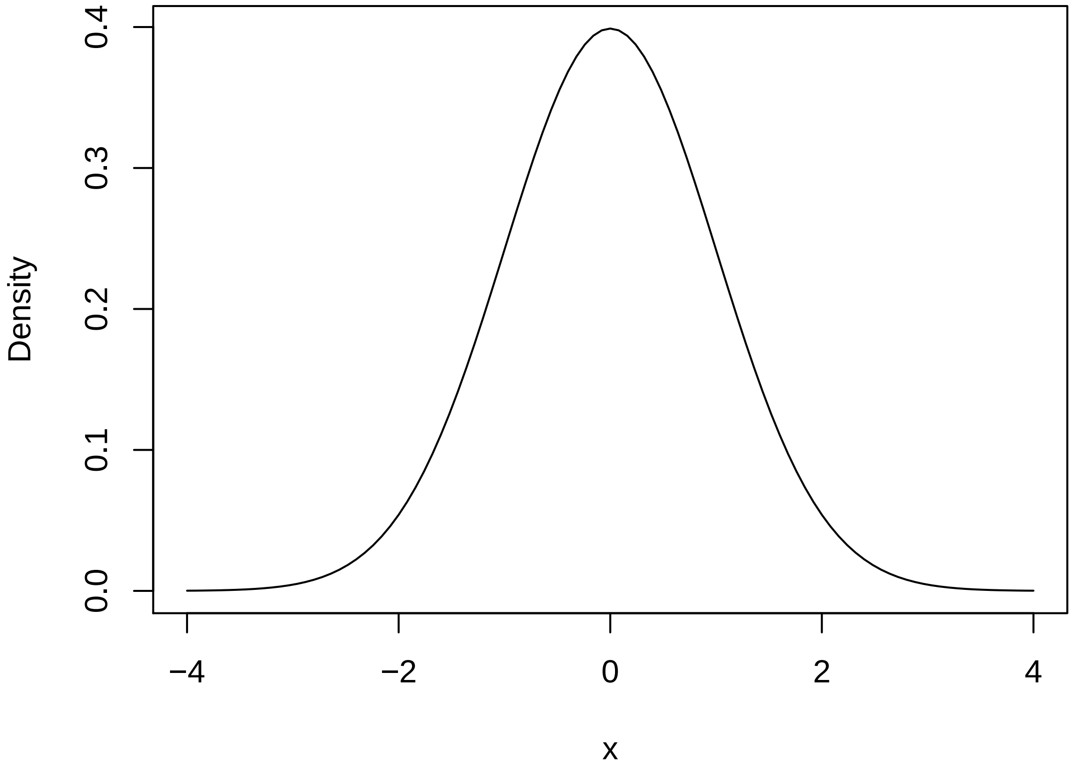
- We have some data of sample size 100,000, if we draw Density Histogram and make the breakpoints finer and finer...



-

From Histogram to Density

- We will end up having the so-called density curve.



-

PDF

- The probability density function (pdf) of a continuous rv X is a function $f(x)$ such that for any two numbers a and b with $a \leq b$,

  $$\mathrm{P}(a \leq \mathrm{X} \leq b) = \int_a^b f(x)dx$$

  The graph of $f(x)$ is often referred to as the *density curve*.

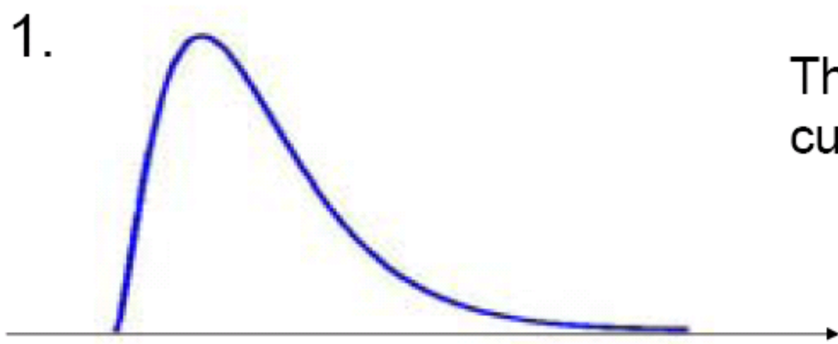
- This means the area under the density curve represents probability!

- Note that $0 \leq f(x)$ for all x.

- $f(x)dx$ can be treated as P(X=x)!

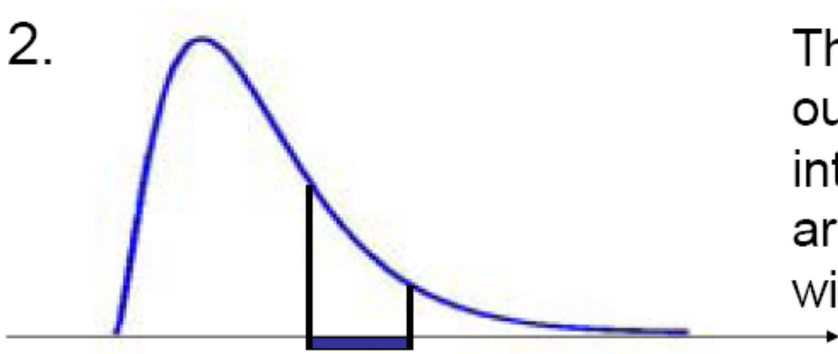
Properties of PDF

1.



The total area under the curve must equal 1.

2.



The probability that the outcome lies in a specific interval is given by the area under the curve within that interval.

Uniform Distribution

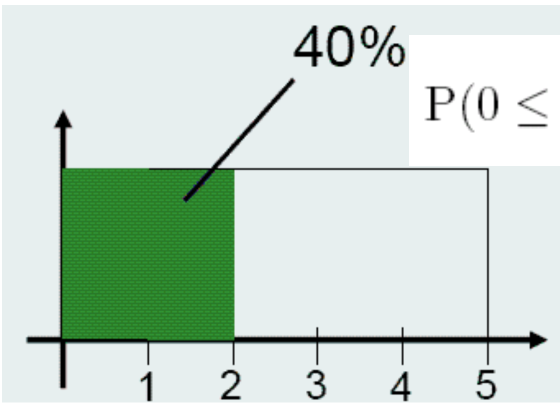
- A continuous rv X is said to have a uniform distribution on the interval [*A*, *B*] if the pdf of X is

$$f(x; A, B) = \begin{cases} \frac{1}{B-A} & A \leq x \leq B \\ 0 & \text{otherwise} \end{cases}$$

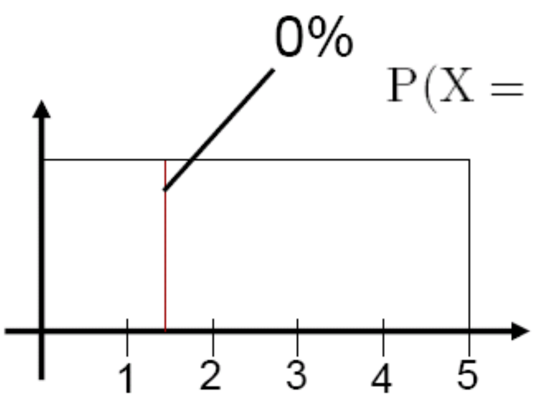
- Verify that this is a proper pdf.
  1. *f*(x) ≥ 0 for all *x*.
  2. Area under *f*(x) should be equal to 1.

Example

Ex. Suppose a bus arrives equally likely at any time between 7:00 – 7:05 AM. What is the probability it arrives sometime between 7:00 – 7:02 AM?



$$P(0 \leq X \leq 2) = \int_0^2 \frac{1}{5} dx = \frac{2}{5}$$



$$P(X = c) = \lim_{\epsilon \to 0} P(c - \epsilon \leq X \leq c + \epsilon) = \lim_{\epsilon \to 0} \int_{c-\epsilon}^{c+\epsilon} \frac{1}{B - A} dx = 0$$

The CDF

- Although the idea of pmd does not extend to the continuous rv's, the idea of cdf still works.

- The cumulative distribution function (cdf) $F(x)$ for a continuous rv X is defined for every number x by

$$F(x) = \mathrm{P}(\mathrm{X} \leq x) = \int_{-\infty}^{x} f(y)dy$$

- $F(x)$ is in fact the probability that a rv X is smaller than x. $F(x)$ increases smoothly as x increases. $F(-\infty) = 0$, $F(+\infty) = 1$.

- It is easy to compute probabilities using $F(x)$.
  - $\mathrm{P}(\mathrm{X} > a) = 1 - F(a)$
  - $\mathrm{P}(a \leq \mathrm{X} \leq b) = F(b) - F(a)$

pdf from cdf

- If X is a continuous rv with pdf $f(x)$ and cdf $F(x)$, then at every x at which the derivative $F'(x)$ exists, $F'(x) = f(x)$. $f(x)$ is often a smooth curve, which is the probability density function (pdf) of X.

- Let p be a number between 0 and 1. The (100p)th percentile (quantile) of the distribution of a continuous rv X, denoted by $\eta(p)$, is defined by

$$p = F(\eta(p)) = \int_{-\infty}^{\eta(p)} f(y)dy$$

- The median of a continuous distribution, denoted by $\tilde{\mu}$, is the 50th percentile, so $\tilde{\mu}$ satisfies .5 = F($\tilde{\mu}$). That is, half the area under the density curve is to the left of $\tilde{\mu}$ and half is to the right of $\tilde{\mu}$.

Expected Values

- Notice that the pdf *f*(*x*) of a continuous distribution is actually playing the role of pmf *p*(*x*) of a discrete distribution.

- Recall that the expected value of a discrete distribution is calculated by

$$\mu_X = \mathrm{E}(\mathrm{X}) = \sum_{x \in D} x \cdot p(x)$$

- Therefore, similarly we can define the expected value of a continuous distribution by

$$\mu_X = \mathrm{E}(\mathrm{X}) = \int_{-\infty}^{\infty} x \cdot f(x) dx$$

- Take advantage of the *symmetry* of particular distributions, when calculating expectations.

Variance

- With a similar argument as in the discrete case, we can also define the expectation of a function of a continuous rv as well as the variance of a continuous rv.

- Proposition: if X is a continuous rv with pdf *f*(*x*) and *h*(X) is any function of X, then

$$\mathrm{E}[h(\mathrm{X})] = \int_{-\infty}^{\infty} h(x) \cdot f(x) dx$$

- As a special case of the above proposition, the variance of X is defined by

$$\sigma_X^2 = \mathrm{Var}(\mathrm{X}) = \mathrm{E}(\mathrm{X} - \mathrm{E}(X))^2 = \int_{-\infty}^{\infty} (x - \mu_X)^2 \cdot f(x) dx$$

The standard deviation (SD) of X is $\sigma_X = \sqrt{\mathrm{Var}(\mathrm{X})}$.

Properties

- Some properties of mean and variance hold in the continuous case in a similar way as in the discrete case.

- For example, under linear transformation of X, we have

1. $E(aX+b) = aE(X) + b$
2. $Var(aX+b) = a^2Var(X)$

- Exercise: prove the above formulas rigorously!

Uniform RV

- We call a uniform rv U a standard uniform, if and only if U ~ uniform on [0,1]

- For a standard uniform rv U, we can easily calculate,

$$\mathrm{E}(\mathrm{U}) = \int_0^1 x \cdot 1 dx = \frac{1}{2}$$

$$\mathrm{E}(\mathrm{U}^2) = \int_0^1 x^2 \cdot 1 dx = \frac{1}{3}$$

$$\mathrm{Var}(\mathrm{U}) = \mathrm{E}(\mathrm{U}^2) - [\mathrm{E}(\mathrm{U})]^2 = \frac{1}{12}$$

General Uniform

- Note that a general case of uniform distribution X on [*A, B*] can be treated as a linear transform of a standard uniform, i.e., $X = (B - A)U + A$.

- Proposition:

> If X is a continuous uniform rv on [*A, B*], then
> $E(X) = (B + A)/2$, $Var(X) = (B - A)^2/12$

- R command:

```
dunif(x, min=0, max=1),
punif(q, min=0, max=1),
qunif(p, min=0, max=1).
```