Precept 2: Random Variables

Soc 500: Applied Social Statistics

Ian Lundberg

Princeton University

September 2016

Acknowledgments

I learned probability in courses taught by Joseph Blitzstein, Carl Morris, and Jessica Hwang. These slides draw frequently on examples from their notes and lectures, and from the Blitzstein and Hwang *Introduction to Probability* textbook. You should look at it if you want more examples to build intuition with random variables! Several slides also come from Brandon Stewart.

Logistics

- Reactions to the problem set?
- Solutions will be posted at 9:30
- New problem set is out

Learning Objectives

1. Build intuition with random variables
2. Comfort applying the rules of expectation and variance
3. Review Benford's Law (useful for homework)
4. Conceptual clarity with joint distributions and marginalization
5. Convey that random variables are fun!

Key formulas to review

CDF: $F_X(x) = P(X < x)$

PDF: $f_X(x) = \frac{\partial}{\partial x} F_X(x)$ (this $\neq P(X = x)$, which is 0!)

Definition of expectation: $E(X) = \begin{cases} \sum_x xp(x), & \text{if } x \text{ discrete} \\ \int xf(x)dx, & \text{if } x \text{ continuous} \end{cases}$

LOTUS (Law of the Unconscious Statistician):

$$E(g[X]) = \begin{cases} \sum_x g(x)p(x), & \text{if } x \text{ discrete} \\ \int g(x)f(x)dx, & \text{if } x \text{ continuous} \end{cases}$$

Adam's law (iterated expectation): $E(Y) = E(E[Y \mid X])$

Evve's law (total variance): $V(Y) = E(V[Y \mid X]) + V(E[Y \mid X])$

Key formulas to review

Variance of sums:

$$V(aX + bY) = a^2X + b^2Y + 2abCov(X, Y)$$

$$V(aX - bY) = a^2x + b^2Y - 2abCov(X, Y)$$

Definition of variance:

$$\begin{aligned} V(X) &= E\left[(X - E[X])^2\right] \\ &= E\left(X^2 - [E(X)]^2\right) \end{aligned}$$

Definition of covariance:

$$\begin{aligned} Cov(X, Y) &= E\left[(X - E[X])\left(Y - E[Y]\right)\right] \\ &= E\left(XY - E[X]E[Y]\right) \end{aligned}$$

Definition of correlation: $Cor(X, Y) = \frac{Cov(X,Y)}{SD(X)SD(Y)}$

What is a random variable?

Definition

A **random variable** is a mapping from the sample space to the real line.

Example:

- Y is the income of a random person in a country
- Ω = sample space = incomes of all people in the country
- Y is a function mapping the chosen person to an income
- Randomness is in the person who was chosen

What is this $\sim$ sign?

- Equality in distribution
- Does not imply equality in values
- Can we think of two random variables with the same distribution, which are not necessarily equal?
  - Two coin flips
- ...which are almost surely not equal?

$$Z_1 \sim Normal(0, 1)$$

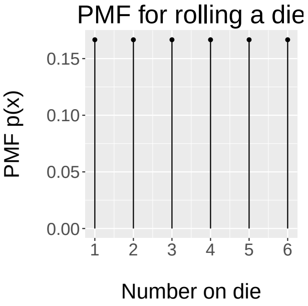
$$Z_2 = -Z_1$$

Probability mass function (PMF)

When I roll a die, what is the PMF?

$$P(X = 1) = 1/6, ..., P(X = 6) = 1/6$$

$$P(X = x) = \begin{cases} 1/6, & \text{if } x \in \{1, ..., 6\} \\ 0, & \text{otherwise} \end{cases}$$



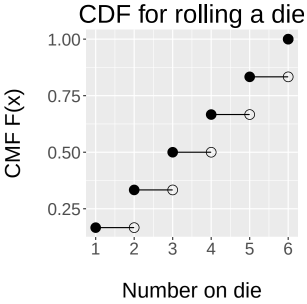
How we made that PMF figure

```
ggplot(data = data.frame(x = 1:6,
                         y = rep(1/6, 6),
                         yend = rep(0,6)),
       aes(x = x, y = y, xend = x, yend = yend)) +
  geom_point() +
  geom_segment() +
  scale_x_continuous(name="\nNumber on die", breaks=1:6) +
  scale_y_continuous(name="PMF p(x)\n") +
  ggtitle("PMF for rolling a die") +
  theme(text=element_text(size=20)) +
  ggsave("DiePMF.pdf",
         height=4,
         width=4)
```

When I roll a die, what is the CDF?

$$P(X \leq 1) = 1/6, \ldots, P(X \leq 6) = 6/6$$

$$P(X \leq x) = \begin{cases} 0, & x < 1 \\ \lfloor x \rfloor / 6, & x \in [1, 6] \\ 1, & x > 6 \end{cases}$$



How we made that CDF figure

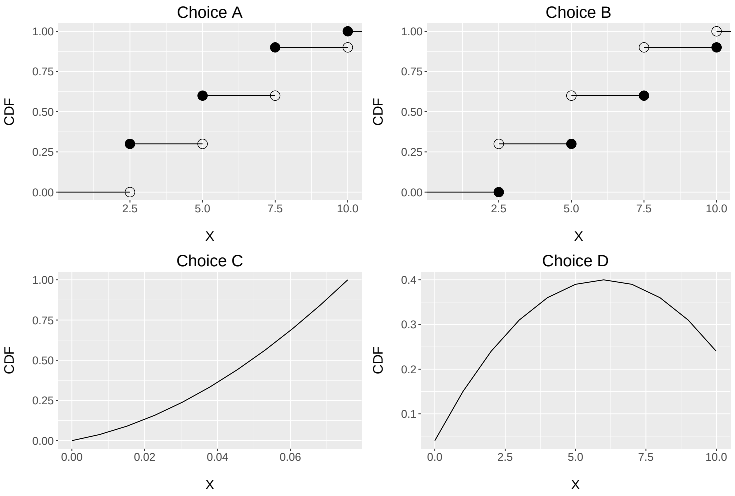
```
ggplot(data = data.frame(x = 1:6,
                         y = (1:6)/6,
                         xend = c(2:6,6))) +
  geom_point(aes(x = x, y = y), size = 4) +
  geom_segment(aes(x = x, xend = xend, y = y, yend = y)) +
  geom_point(aes(x = xend, y = y), shape = 1, size = 4) +
  scale_x_continuous(name="\nNumber on die",
                     breaks=1:6) +
  scale_y_continuous(name="CDF F(x)\n") +
  ggtitle("CDF for rolling a die") +
  theme(text=element_text(size=20)) +
  ggsave("DieCDF.pdf",
         height=4,
         width=4)
```

Which is a proper CDF? (check all that apply)

Respond at **PollEv.com/ianlundberg444**

Text **IANLUNDBERG444** to **37607** once to join, then text your message

Answers to this poll are anonymous



Properties of the CDF

- Non-decreasing
- Right continuous
- $F(x) \to 0$ as $x \to -\infty$
- $F(x) \to 1$ as $x \to \infty$

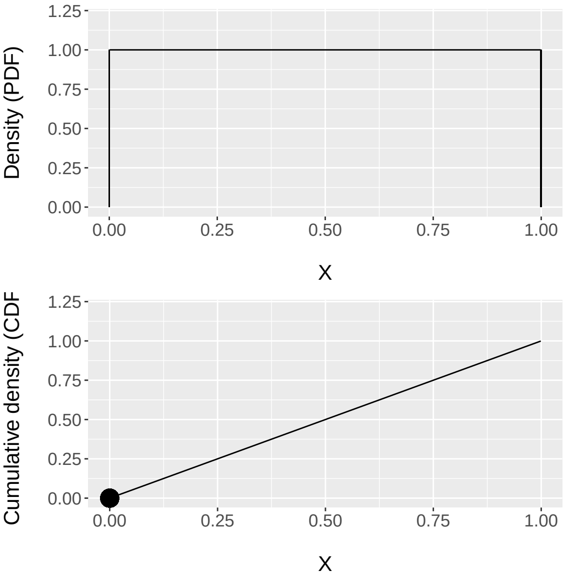
Continuous random variables

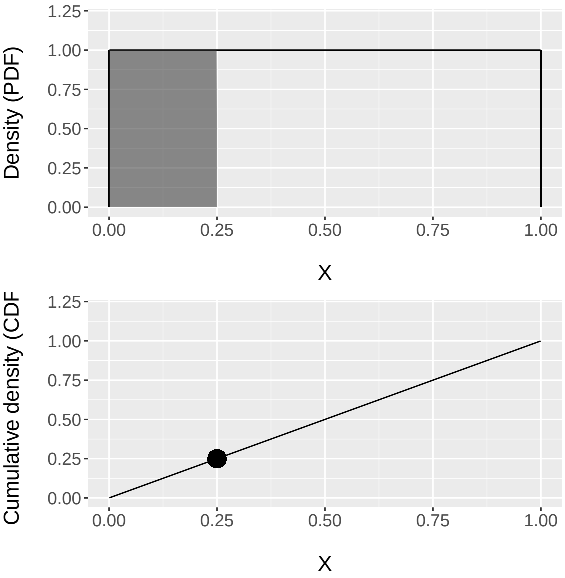
Suppose it is Lawnparties, and a very drunk Princetonian spins around ten times before throwing darts at a wall. Suppose the sides of the wall are marked 0 and 1. Ignoring the vertical position of the darts, the horizontal position of the darts might be distributed **uniformly** over the interval.

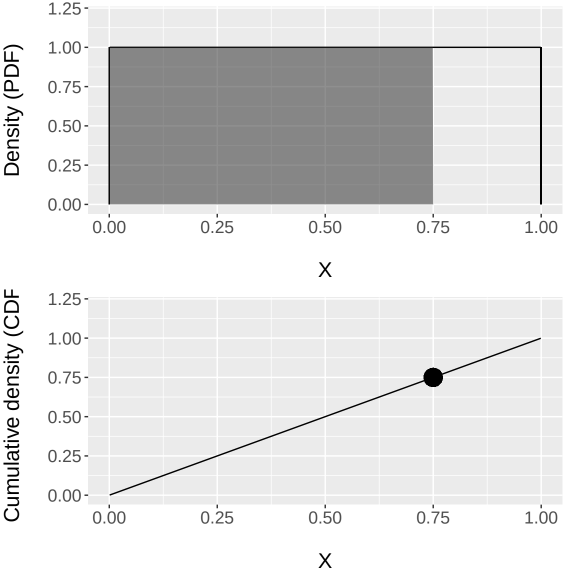
$$U \sim Uniform(0, 1)$$

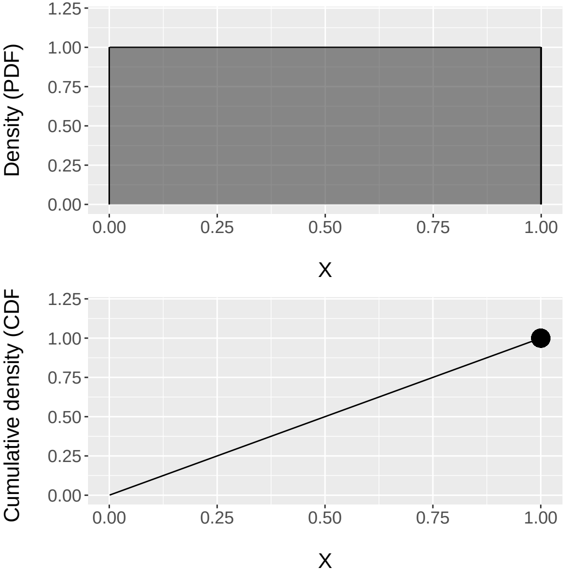
The CDF of the uniform is

$$F(x) = x$$

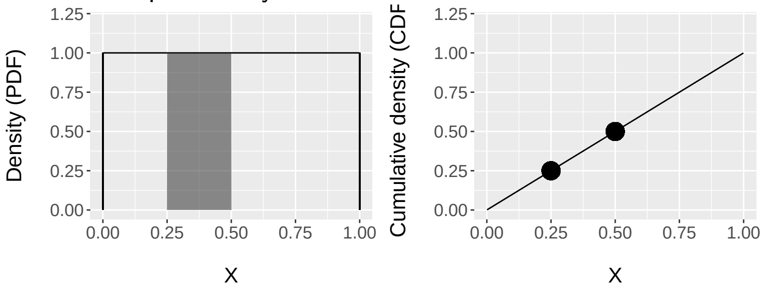

Density (PDF)
1.25
1.00
0.75
0.50
0.25
0.00
0.00
0.25
0.50
0.75
1.00
X
Cumulative density (CDF
1.25
1.00
0.75
0.50
0.25
0.00
0.00
0.25
0.50
0.75
1.00
X


Density (PDF)
1.25
1.00
0.75
0.50
0.25
0.00
0.00 0.25 0.50 0.75 1.00
X
Cumulative density (CDF
1.25
1.00
0.75
0.50
0.25
0.00
0.00 0.25 0.50 0.75 1.00
X


Density (PDF)
1.25
1.00
0.75
0.50
0.25
0.00
0.00 0.25 0.50 0.75 1.00
X
Cumulative density (CDF
1.25
1.00
0.75
0.50
0.25
0.00
0.00 0.25 0.50 0.75 1.00
X


Density (PDF)
1.25
1.00
0.75
0.50
0.25
0.00
0.00 0.25 0.50 0.75 1.00
X
Cumulative density (CDF
1.25
1.00
0.75
0.50
0.25
0.00
0.00 0.25 0.50 0.75 1.00
X

What is the probability that a dart lands between 0.25 and 0.5?



$$F(.5) - F(.25) = .25$$

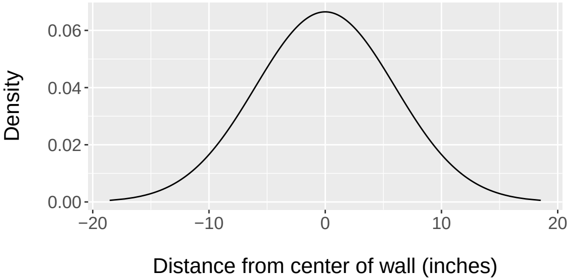
The `punif()` command in `R` is the uniform cumulative distribution function: `punif(.5) - punif(.25)`

Now let's suppose we have someone better at throwing darts, so we're going to measure how far they are from the center of the wall in inches. In this case, perhaps the darts will be distributed **normally** around the center of the wall.

$$X \sim N(\mu = 0, \sigma = 12)$$

$$X \sim N(\mu = 0, \sigma = 12)$$

What is the probability of getting a bullseye ($X = 0$)?



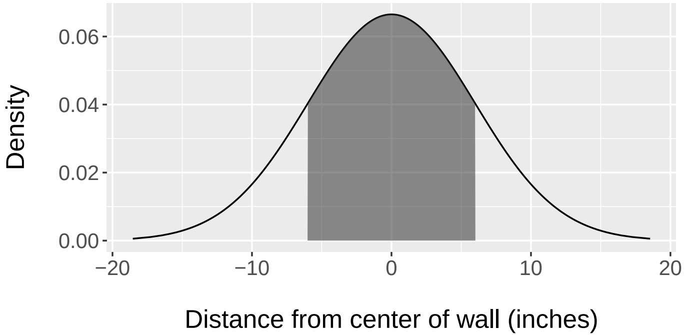
Respond at **PollEv.com/ianlundberg444**

Text **IANLUNDBERG444** to **37607** once to join, then text your message

Answers to this poll are anonymous

...we'll come back to that.

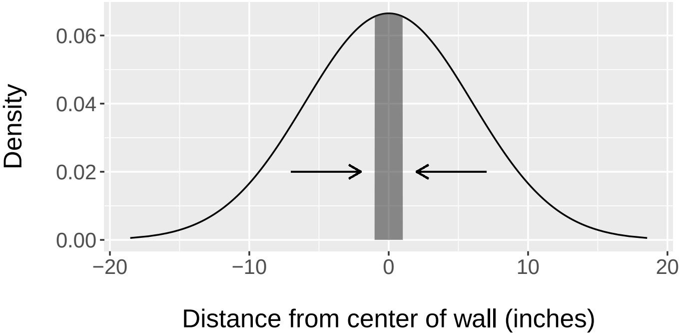
How would we calculate the probability that a dart lands within 6 inches of the center of the wall?



$$P(X \in (-6, 6)) = P(X < 6) - P(X < -6) = \Phi(6) - \Phi(-6)$$

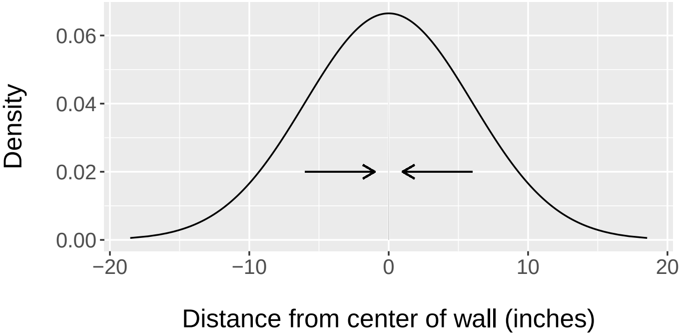
```
pnorm(6, mean = 0, sd = 12) - pnorm(-6, mean = 0, sd = 12)
```

One inch?



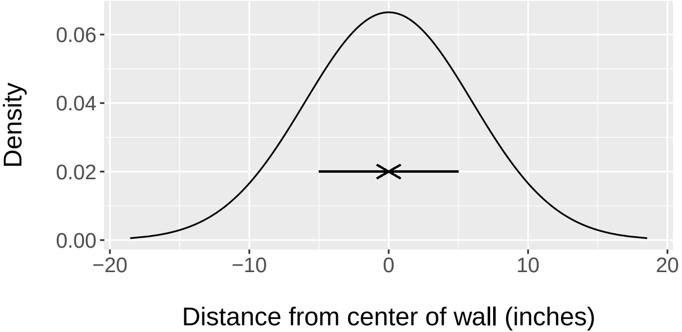
$$P(X \in (-1, 1)) = 0.0664135$$

$1/100$th of an inch?



$$P(X \in (-.01, .01)) = 0.0006649037$$

A perfect bullseye?



$$P(X = 0) = 0$$

The probability that a continuous variable takes on any particular discrete value is **0**!

Useful R functions

Let's type `?dnorm` into our R consoles.
There are 4 functions for the normal distribution:

- `dnorm` gives the PDF
- `pnorm` gives the CDF
- `qnorm` is $F^{-1}()$: it gives quantiles
- `rnorm` generates random draws

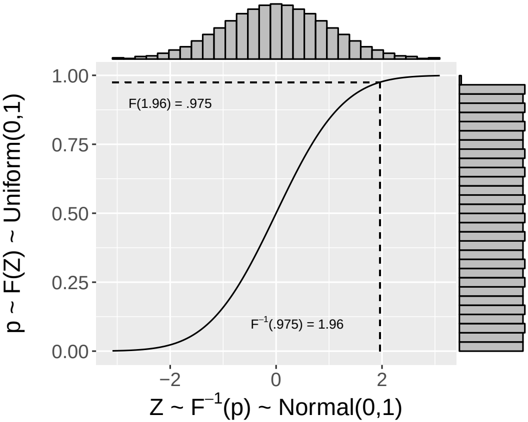
Question for the class: Can we recreate the `rnorm` using `qnorm` and `runif`?

Universality of the Uniform

aka Probability Integral Transform, PIT

Theorem

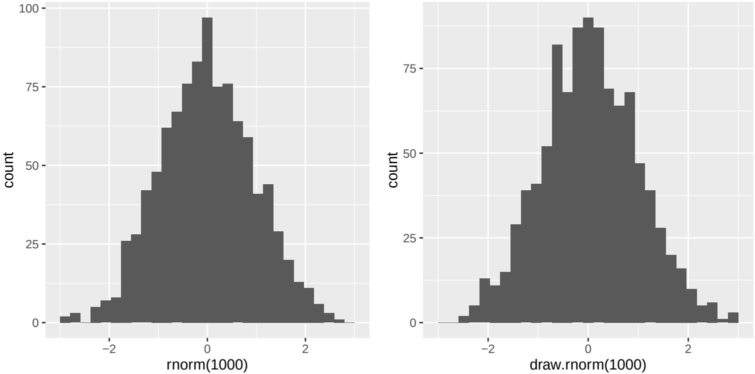
- *Regardless of the distribution of* X, $F(X) \sim \mathit{Uniform}(0,1)$
- *For a r.v.* X *with CDF* F *and a Uniform r.v.* U, $F^{-1}(U) \sim X$



Self-written rnorm

```
draw.rnorm <- function(n, mean = 0, sd = 1) {
  u <- runif(1)
  z <- qnorm(u)
  return(z)
}
qplot(x=draw.rnorm(1000), geom="histogram",
      bins = 30, xlim = c(-3,3))
qplot(x=rnorm(1000), geom="histogram",
      bins = 30, xlim=c(-3,3))
```

Did we succeed?



Definition of expectation

$$E(X) = \begin{cases} \sum xp(x), & \text{if } X \text{ is discrete} \\ \int xf(x)dx, & \text{if } X \text{ is continuous} \end{cases}$$

If you can write down the PDF or PMF, you can calculate the expected value.

Example: High jumpers

Athletes compete one by one in a high jump competition. Let X_j be how high the jth person jumped (in meters). Let m be the median of the distribution of X_j. Let N be the number of jumpers before the first one who successfully jumps higher than m meters, not including that jumper. Find the mean of N.
What do we need? The PMF of N.
What do we know? For each jumper i, $P(X_i < m) = 0.5$
Can we write down the PMF of N?

$$P(N = 0) = P(\text{first jumper clears}) = 0.5$$

$$P(N = 1) = P(\text{first jumper doesn't clear, and second does}) = (0.5)0.5$$

$$P(N = 2) = P(\text{first 2 jumpers don't clear, and third does}) = (0.5^2)0.5$$

High jumpers (continued)

$$
\begin{aligned}
E(N) &= 0P(N=0) + 1P(N=1) + 2P(N=2) + \ldots \\
&= 0(.5) + 1(.5).5 + 2(.5^2).5 + \ldots \\
&= \sum_{x=0}^{\infty} x\left(.5^{x+1}\right) \\
&= 1 \text{ by plugging in to WolframAlpha.com}
\end{aligned}
$$

The expected number of high jumpers to attempt before one clears the median jump height is 1.

Linearity of Expectation: Questions

Suppose $E(X) = \mu_x$ and $E(Y) = \mu_y$. Solve the following:

1. $E(X + Y) =$
2. $E(aX) =$
3. $E(aX + bY) =$
4. If X and Y are independent, what is $E\left(\frac{Y}{2} \mid X < 4\right)$?

Linearity of Expectation: Answers

Suppose $E(X) = \mu_x$ and $E(Y) = \mu_y$. Solve the following:

1. $E(X + Y) = E(X) + E(Y) = \mu_x + \mu_y$
2. $E(aX) = aE(X) = a\mu_x$
3. $E(aX + bY) = aE(X) + bE(Y) = a\mu_x + b\mu_y$
4. If X and Y are independent, what is $E\left(\frac{Y}{2} \mid X < 4\right)$?

Answer:
$E(\frac{Y}{2} \mid X < 4) = E\left(\frac{Y}{2}\right)$ (since independent) $= \frac{1}{2}E(Y) = \frac{\mu_y}{2}$

Linearity of Expectation

Suppose I flip a coin 20 times. What is the expected value of the number of heads?
for $i \in 1, \ldots, 20$,

$$X_i = \begin{cases} 1, & \text{if heads} \\ 0, & \text{if tails} \end{cases}$$

$$Y = (\#\text{ heads}) = X_1 + X_2 + ...X_{20}$$

$$E(Y) = E\left(\sum_{i=1}^{20} X_i\right)$$

$$= \sum_{i=1}^{20} E(X_i)$$

$$= \sum_{i=1}^{20} 0.5 = 10$$

Law of Iterated Expectation (Adam's Law)

Theorem

$$E(E[Y \mid X]) = E(Y)$$

When calculating an expected value is hard, but would be easy if you knew something else, **you can condition on that something else!**

An example: Parking at the beach

Parking tickets at the beach

Suppose you go to the beach in Belmar, NJ, and you park in the first row of spaces where there are meters, but you forget to pay the meter. Suppose the probability that you get a parking ticket (C for caught) is proportional to the time spent there (T), with the probability increasing with time according to the function

$$P(C \mid T) = \frac{1}{5}\left(T - \frac{1}{9}T^2\right)$$

Suppose the expected parking ticket is \$80 if you are caught, and when you go to the beach you stay 3 hours on average with a variance of 2 hours. What is your average parking ticket (F for fine) when you forget to pay the meter?

What do we know? $E(F \mid C=1)=80, E(F \mid C=0)=0, P(C \mid T)=\frac{1}{5}\left(T-\frac{1}{9}T^2\right), E(T)=3, V(T)=2$
What do we want? $E(F)$

$$E(F)=E[E(F \mid C)]=E[80C+0(1-C)]=80E(C)$$

$$E(C)=E(E[C \mid T])=E\left(\frac{1}{5}\left(T-\frac{1}{9}T^2\right)\right)$$

$$=\frac{1}{5}E(T)-\frac{1}{45}E(T^2)$$

$$V(T)=E(T^2)-[E(T)]^2 \rightarrow E(T^2)=V(T)+[E(T)]^2=2+3^2=11$$

$$E(C)=\frac{1}{5}(3)-\frac{1}{45}(11)=\frac{27}{45}$$

$$E(F)=80E(C)=80*\frac{16}{45}=\$28.44$$

Real advice

Variance is not a linear operator

Variance definition and rules

$$V(X) = E(X - E[x])^2$$

$$V(X) = E(X^2) - (E[x])^2$$

$$V(aX + bY) = a^2 V(X) + b^2 V(Y) + 2abCov(X, Y)$$

Four problems: (assume $V(X) = \sigma_x^2$, $V(Y) = \sigma_y^2$, and $Cov(X, Y) = \rho$)

1. $V(5X) = 25V(X) = 25\sigma_x^2$
2. $V(X + 4) = V(X) = \sigma_x^2$
3. $V(3X - 2Y) = 3^2 V(X) + 2^2 V(Y) - 2 * 3 * 2Cov(X, Y) =$ $9\sigma_x^2 + 4\sigma_y^2 - 12\rho$
4. $V(X \mid X^2) = E(X^2 \mid X^2) - [E(X \mid X^2)]^2 = X^2 - 0 = X^2$

(assume for 4 that the distribution of X is symmetric around 0)

Law of Total Variance (Evve's Law)

Theorem

$$V(Y) = E(V[Y \mid X]) + V(E[Y \mid X])$$

Intuition: Suppose Y is income and X is education. This says that the total variance in income can be divided into the expected amount of **variance within categories** of education $E(V[Y \mid X])$ and the variance in expected incomes **between categories** $V(E[Y \mid X])$.

We might think of $E(V[Y \mid X])$ as the variance in income that is not explained by education categories. More to come later in the semester!

Theorem

$$V(Y) = E(V[Y \mid X]) + V(E[Y \mid X])$$

Our fine example again (pun not intended): $F =$ fine, $C =$ whether caught.

$$V(F) = E(V(F \mid C)) + V(E(F \mid C))$$

$$= E(0) + V(80C)$$

$$= 0 + 80^2 V(C)$$

What is $V(C)$?

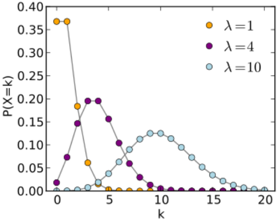
$$V(C) = E(V(C \mid T)) + V(E(C \mid T))$$

This gets to ugly math, but you see the point. Conditioning makes the problem manageable!

Poisson mean and variance in R: Useful for homework!

A **Poisson** random variable

- captures counts of events
- is parameterized by a rate of events, λ
- both the mean and the variance are λ



Let's verify the mean and variance by simulation in `R`.

Poisson mean and variance in R: Useful for homework!

```
## Initialize Lambda vector
lambdas <- c(0.5, 1, 2)
## Initialize table to store results
pois_mean_var <- matrix(data = NA, nrow = 3, ncol = 2)
rownames(pois_mean_var) <- as.character(lambdas)
colnames(pois_mean_var) <- c("Mean", "Variance")
## Set a random seed
set.seed(08544)
for (lambda in lambdas) {
  pois_sim <- rpois(2000, lambda = lambda)
  ## Store simulated values in pois_mean_var
  pois_mean_var[as.character(lambda), ] <-
    c(mean(pois_sim), var(pois_sim))
}
```

Poisson example: Melting data

```
pois_mean_var
      Mean  Variance
0.5 0.4990 0.4992486
1   1.0315 1.0000078
2   1.9775 1.9659767
```

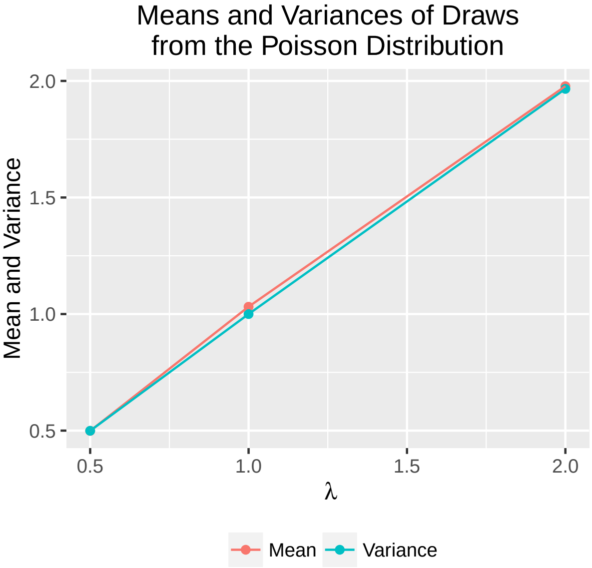
We can use the `reshape2` package's `melt` function to change the shape of that data.

```
pois_mean_var_df <- melt(pois_mean_var)
  Var1     Var2     value
1  0.5     Mean 0.4990000
2  1.0     Mean 1.0315000
3  2.0     Mean 1.9775000
6  0.5 Variance 0.4992486
7  1.0 Variance 1.0000078
8  2.0 Variance 1.9659767
```

Poisson example: easy ggplot()!

```
ggplot(pois_mean_var_df,
       aes(x = Var1,
           y = value,
           color = Var2)) +
  geom_point() +
  geom_line() +
  labs(title = "Means and Variances of Draws\n
                from the Poisson Distribution",
       x = expression(~lambda), y = "Mean and Variance") +
  theme(legend.position = "bottom", legend.title=element_blank())
```

Poisson example: easy ggplot()!



Covariance

Definition

The **covariance** of X and Y is defined as

$$Cov(X, Y) = E\left[(X - E[X])(Y - E[Y])\right]$$

Intuition: If X and Y tend to be higher or lower than expected at the same time, then covariance is positive.
This formula can be simplified to this alternative:

$$Cov(X, Y) = E(XY) - E(X)E(Y)$$

Intuition: Covariance is farther from 0 when it's more true that $E(XY)$ and $E(X)E(Y)$ give different things. Can we show that under independence the covariance is 0?

What Ian looks like at the start of a marthon

...and 26.2 miles later

Marathon times: Practice

	X: Time on first half (hrs)	
Y: Time on second half (hrs)	2	3
2	0.20	0.25
3	0.30	0.25

Table: Distribution of marathon times

1. Find $E(Y \mid X)$ for each possible X
2. Find $E(Y)$
3. Find $E(X)$
4. Find $E(XY)$
5. Find the covariance of X and Y

Benford's Law[1]

1. Law describing the (non-uniform) distribution of leading digits in many real-life sources of data.
2. Found in election results, populations of cities, stock market prices, word frequencies, death rates, street addresses and anything with a power law.
3. Often used for fraud detection.

[1] Benford's Law slides deeply indebted to Brandon Stewart and previous preceptors.

Benford's Law

For base 10 counting systems, Benford's Law states that the leading digit has the probability distribution:

$$P(d) = \log_{10}(d+1) - \log_{10}(d)$$
$$= \log_{10}(1 + \frac{1}{d})$$

Originally discovered by Simon Newcomb in 1881 but then restated and named after physicist Frank Benford in 1938.

Benford's Law

Benford's Law: Applications

- Accounting Fraud (Nigrini, 1999)
- Campaign Finance Fraud (Cho and Gaines, 2007)
- Polling Fraud (Grebner and Weissman 2010)
- Iranian Elections (Beber and Scacco 2009)

Benford's Law is admissible evidence of fraud in U.S. court!

Benford's Law: Data Rules[2]

Benford's Law works best under the following conditions:

- Numbers that result from combinations of other numbers (e.g. quantity times price)
- Individual transactions like sales or data
- Large datasets (these are asymptotic properties!)
- Positive skew with mean greater than the median

It doesn't work as well in situations where:

- Numbers are assigned (check numbers, invoice numbers etc.)
- There are procedural or psychological thresholds

Question: Why wouldn't this work with parking fines?

---

[2] These come directly from Cho and Gaines (2007) formulation of the guidelines in Durtschi, Hillison and Pacini(2004)

Benford's Law: Analytic Practice

Let's practice our analytic skills by looking at the expectation and variance of first digits under Benford's Law:
How would we write the expectation of the first digit?

$$P(d) = \log_{10}(1 + \frac{1}{d})$$

Benford's Law: Analytic Practice

Let's practice our analytic skills by looking at the expectation and variance of first digits under Benford's Law:
How would we write the expectation of the first digit?

$$
\begin{aligned}
P(d) &= \log_{10}(1 + \frac{1}{d}) \\
E(D) &= \sum_{i=1}^{9} \log_{10}(1 + \frac{1}{i}) * i \\
&= 3.44
\end{aligned}
$$

Benford's Law: Analytic Practice

Let's practice our analytic skills by looking at the expectation and variance of first digits under Benford's Law:
How would we write the expectation of the first digit?

$$P(d) = \log_{10}(1 + \frac{1}{d})$$
$$E(d) = \sum_{i=1}^{n} \log_{10}(1 + \frac{1}{i}) * i$$

How would you write the variance?

Benford's Law: Analytic Practice

Let's practice our analytic skills by looking at the expectation and variance of first digits under Benford's Law:
How would we write the expectation of the first digit?

$$
\begin{aligned}
P(d) &= \log_{10}(1 + \frac{1}{d}) \\
E(D) &= \sum_{i=1}^{9} \log_{10}(1 + \frac{1}{i}) * i \\
&= 3.44
\end{aligned}
$$

How would you write the variance?

$$V(D) = \sum_{i=1}^{9} \log_{10}(1 + \frac{1}{i})(i - 3.44)^2$$

Benford's Law: Computational Practice

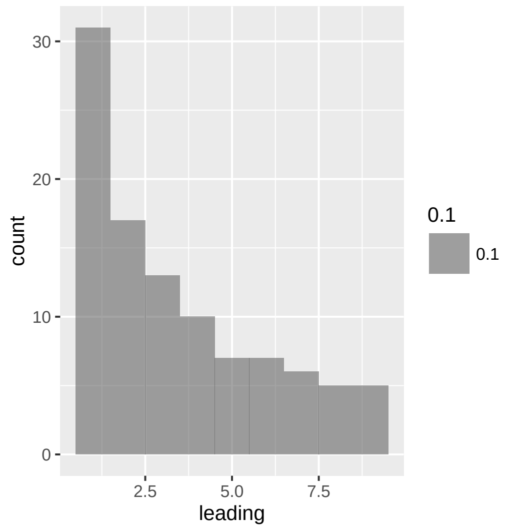
Let's use R to generate the first 100 powers of 2

```
x <- sapply(0:100, function(x) 2^x)
```

Extract the leading digit from those

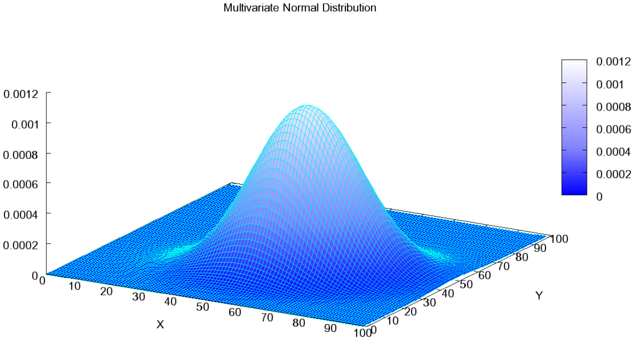
```
extract.lead <- function(x.case) {
  lead <- as.numeric(strsplit(as.character(x.case),"")[[1]])[1]
}
leading <- sapply(x, extract.lead)
qplot(leading, geom="histogram", binwidth=1, alpha=.1)
```

Benford's Law: Computational Practice

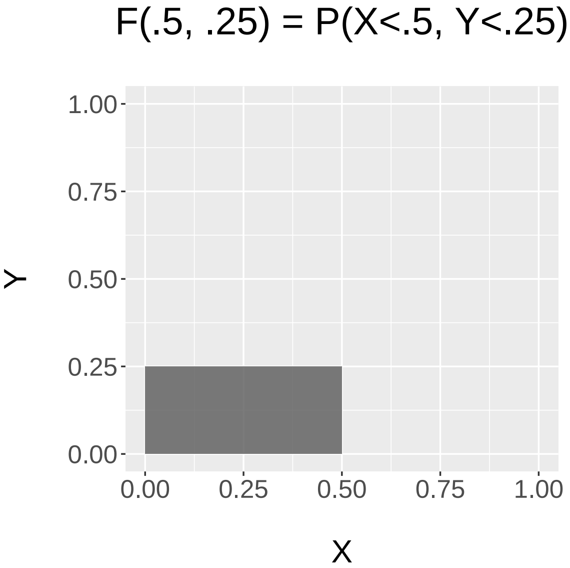


Joint Distributions

The joint distribution of X and Y is defined by a **joint PDF** $f(x, y)$, or equivalently by a **joint CDF** $F(x, y)$.



Join CDF visualization



CDF practice problem 1

Modified from Blitzstein and Morris

Suppose $a < b$, where a and b are constants (for concreteness, you could imagine $a = 3$ and $b = 5$). For some distribution with PDF f and CDF F, which of the following must be true?

1. $f(a) < f(b)$
2. $F(a) < F(b)$
3. $F(a) \leq F(b)$

Joint CDF practice problem 2

Modified from Blitzstein and Morris

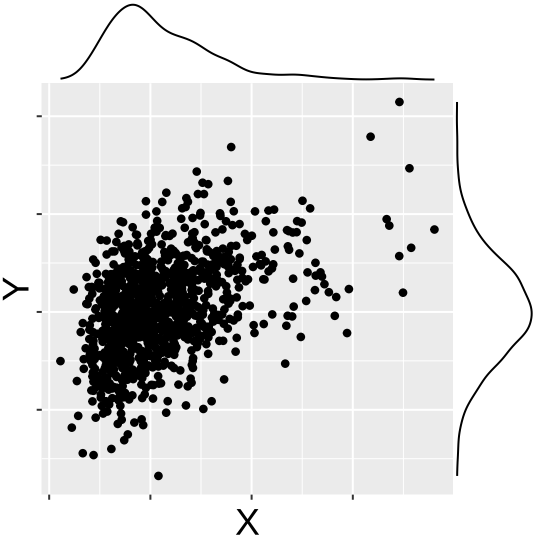
Suppose $a_1 < b_1$ and $a_2 < b_2$. Show that
$F(b_1, b_2) - F(a_1, b_2) + F(b_1, a_2) - F(a_1, a_2) \geq 0$.

$$= [F(b_1, b_2) - F(a_1, b_2)] + [F(b_1, a_2) - F(a_1, a_2)]$$

$$= (\text{something } \geq 0) + (\text{something } \geq 0)$$

$$\geq 0$$

Marginalizing



Correlation between sum and difference of dice

Suppose you roll two dice and get numbers X and Y. What is $Cov(X+Y, X-Y)$?

Note

Covariance is a **bilinear** operator, meaning that things distribute like multiplication.

$$Cov(aX, Y) = aCov(X, Y)$$

$$Cov(aX, bY) = abCov(X, Y)$$

$$Cov(X+Y, Z) = Cov(X, Z) + Cov(Y, Z)$$

$$Cov(X+Y, X-Y) = Cov(X, X) - Cov(X, Y) + Cov(X, Y) - Cov(Y, Y)$$

$$= Var(X) - Var(Y) + Cov(X, Y) - Cov(X, Y)$$

$$= 0 + 0 = 0$$

Are $X+Y$ and $X-Y$ independent? Think of an extreme example.

Example: If $X+Y=2$, then $X=1$ and $Y=1$, so we know $X-Y=0$. Knowing $X+Y$ gives us some information about $X-Y$, so the two are not independent.

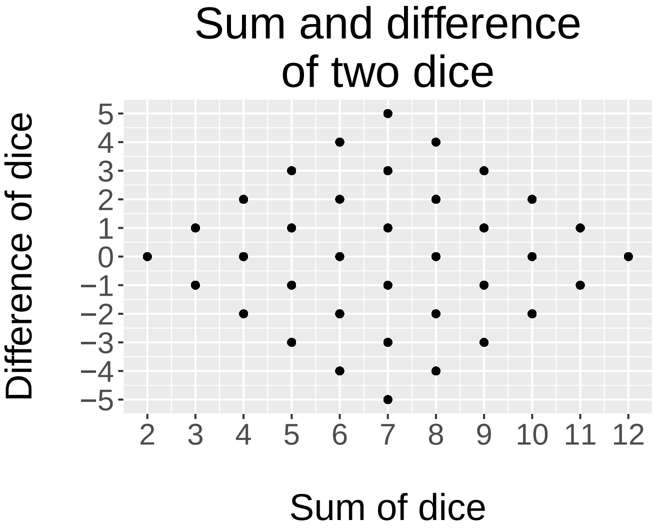
Sum and difference of dice: Simulation

```
draw.sum.diff <- function() {
  x <- sample(1:6,1)
  y <- sample(1:6,1)
  return(c(x+y,x-y))
}
samples <- matrix(nrow=5000,ncol=2)
colnames(samples) <- c("x","y")
set.seed(08544)
for (i in 1:5000) {
  samples[i,] <- draw.sum.diff()
}
```

Sum and difference of dice: Plot

```
ggplot(data.frame(samples), aes(x=x,y=y)) +
  geom_point() +
  scale_x_continuous(breaks=c(2:12),
                     name="\nSum of dice") +
  scale_y_continuous(breaks=c(-5:5),
                     name="Difference of dice\n") +
  ggtitle("Sum and difference\nof two dice") +
  theme(text=element_text(size=20)) +
  ggsave("SumDiff.pdf",
         height=4, width=5)
```

Sum and difference of dice: Plot



Examples to clarify independence

Blitzstein and Hwang, Ch. 3 Exercises

1. Give an example of dependent r.v.s X and Y such that $P(X < Y) = 1$.
2. Can we have independent r.v.s X and Y such that $P(X < Y) = 1$?
3. If X and Y are independent and Y and Z are independent, does this imply X and Z are independent?

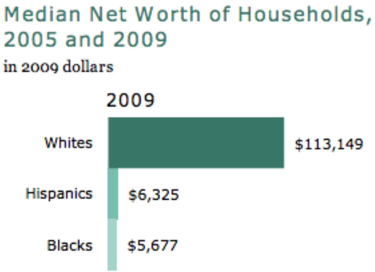
Examples to clarify independence

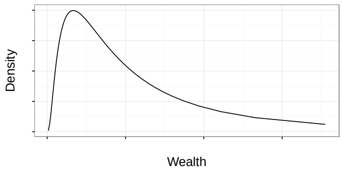
Blitzstein and Hwang, Ch. 3 Exercises

1. Give an example of dependent r.v.s X and Y such that $P(X < Y) = 1$.
    - $Y \sim N(0, 1)$, $X = Y - 1$
2. Can we have independent r.v.s X and Y such that $P(X < Y) = 1$?
    - $X \sim Uniform(0, 1)$, $Y \sim Uniform(2, 3)$
3. If X and Y are independent and Y and Z are independent, does this imply X and Z are independent?
    - No. Consider $X \sim N(0, 1)$, $Y \sim N(0, 1)$, with X and Y independent, and $Z = X$.

Measuring wealth

Wealth inequality in America is dramatic and tightly coupled with race. The figure below shows estimated median net worth by race in 2005 and 2009, from a study by the Pew Research Center (link here). The wealth distribution is also right-skewed (looks something like this simulated distribution).





Question: Would the mean of the wealth distribution be higher? Which report would you prefer?

Wealth, part 2

Suppose wealth is distributed log-normally, such that the log of wealth is normally distributed with mean 8.5 and standard deviation 1. Suppose a survey asks about wealth, but truncates responses above $100,000. Can we find the expected report, both (a) analytically and (b) using `R`? Let Y represent reported wealth and Z represent true wealth.

Analytically:

$$E(Y) = E(Y \mid Z < 100,000)P(Z < 100,000)$$

$$+E(Y \mid Z \geq 100,000)P(Z \geq 100,000)$$

Wealth in R

In R:

```
set.seed(08544)
z <- exp(rnorm(n = 10000, mean = 8.5, sd = 1))
y <- z
y[y>=100000] <- 100000
mean(z)
mean(y)
```

How does median wealth compare to the mean?

```
median(z)
```

The power of conditioning: Coins

Suppose your friend has two coins, one which is fair and one which has a probability of heads of $3/4$. Your friend picks a coin randomly and flips it. What is the probability of heads?

$$P(H) = P(H \mid F)P(F) + P(H \mid F^C)P(F^C)$$

$$= 0.5(.05) + .75(.05) = 0.625$$

Suppose the flip was heads. What is the probability that the coin chosen was fair?

$$P(F \mid H) = \frac{P(H \mid F)P(F)}{P(H)}$$

$$= \frac{0.5(0.5)}{0.625)} = 0.4$$

Exponential: $-\log$ Uniform

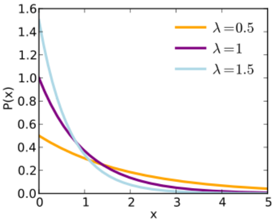
Figure credit: Wikipedia

The Uniform distribution is defined on the interval $(0, 1)$. Suppose we wanted a distribution defined on all positive numbers.

Definition

X follows an **exponential** distribution with rate parameter λ if

$$X \sim -\frac{1}{\lambda} log(U)$$



Exponential: $-\log$ Uniform

The exponential is often used for wait times. For instance, if you're waiting for shooting stars, the time until a star comes might be exponentially distributed.

Key properties:

- Memorylessness: Expected remaining wait time does not depend on the time that has passed
- $E(X) = \frac{1}{\lambda}$
- $V(X) = \frac{1}{\lambda^2}$

Exponential-uniform connection

Suppose $X_1, X_2 \overset{iid}{\sim} Expo(\lambda)$. What is the distribution of $\frac{X_1}{X_1+X_2}$? The proportion of the wait time that is represented by X_1 is uniformly distributed over the interval, so

$$\frac{X_1}{X_1 + X_2} \sim Uniform(0, 1)$$

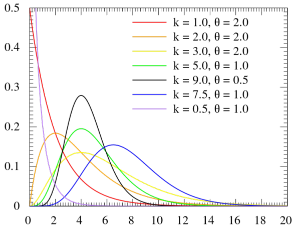
Gamma: Sum of independent Exponentials

Figure credit: Wikipedia

Definition

Suppose we are waiting for a shooting stars, with the time between stars $X_1, \ldots, X_a \stackrel{iid}{\sim} Expo(\lambda)$. The distribution of time until the ath shooting star is

$$G \sim \sum_{i=1}^{a} X_i \sim Gamma(a, \lambda)$$



Gamma: Properties

Properties of the $Gamma(a, \lambda)$ distribution include:

- $E(G) = \frac{a}{\lambda}$
- $V(G) = \frac{a}{\lambda^2}$

Beta: Uniform order statistics

Suppose we draw $U_1, \dots, U_k \sim \textit{Uniform}(0,1)$, and we want to know the distribution of the jth *order statistic*, $U_{(j)}$. Using the Uniform-Exponential connection, we could also think of these $U_{(j)}$ as being the location of the jth Exponential in a series of $k+1$ Exponentials. Thus,

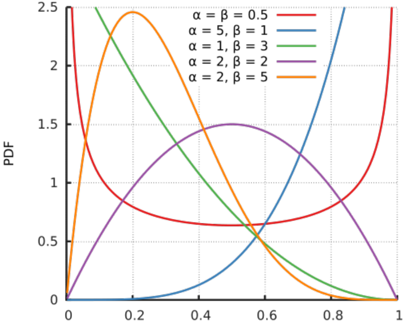
$$U_{(j)} \sim \frac{\sum_{i=1}^{j} X_i}{\sum_{i=1}^{j} X_i + \sum_{i=j+1}^{k+1} X_i} \sim \textit{Beta}(j, k-j+1)$$

This defines the **Beta distribution**.
Can we name the distribution at the top of the fraction?

$$\sim \frac{G_j}{G_j + G_{k-j+1}}$$

What do Betas look like?

Figure credit: Wikipedia



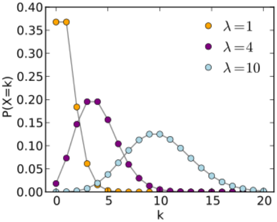
Poisson: Number of Exponential events in a time interval

Figure credit: Wikipedia

Definition

Suppose the time between shooting stars is distributed $X \sim Expo(\lambda)$. Then, the number of shooting stars in an interval of time t is distributed

$$Y_t \sim Poisson(\lambda t)$$



Poisson: Number of Exponential events in a time interval

Properties of the Poisson:

- If $Y \sim Pois(\lambda t)$, then $V(Y) = E(Y) = \lambda t$
- Number of events in disjoint intervals are independent

χ^2_n: A particular Gamma

Definition

We define the **chi-squared distribution** with n degrees of freedom as

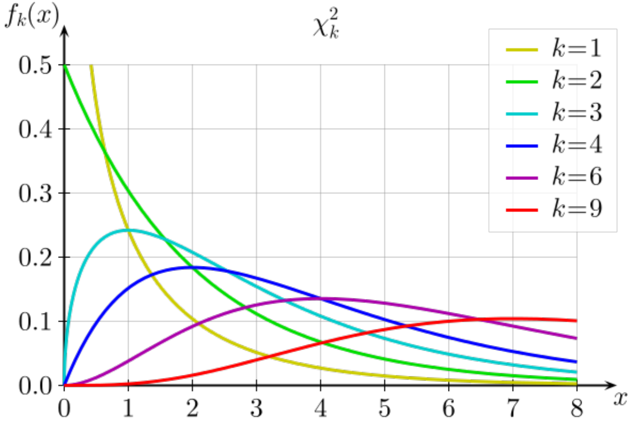
$$\chi^2_n \sim Gamma\left(\frac{n}{2}, \frac{1}{2}\right)$$

More commonly, we think of it as the sum of a series of independent squared Normals, $Z_1, \ldots, Z_n \overset{iid}{\sim} Normal(0,1)$:

$$\chi^2_n \sim \sum_{i=1}^{n} Z_i^2$$

χ_n^2: A particular Gamma

Figure credit: Wikipedia

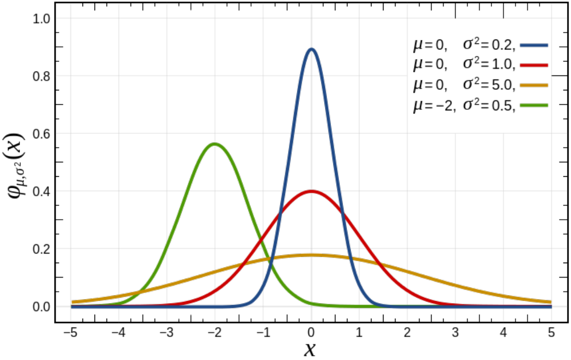


Normal: Square root of χ_1^2 with a random sign

Figure credit: Wikipedia

Definition

Z follows a **Normal** distribution if $Z \sim S\sqrt{\chi_1^2}$, where S is a random sign with equal probability of being 1 or -1.



Normal: An alternate construction

Note: This is far above and beyond what you need to understand for the course!

Box-Muller Representation of the Normal

Let $U_1, U_2 \overset{iid}{\sim} \textit{Uniform}$. Then

$$Z_1 \equiv \sqrt{-2 \log U_2} \cos(2\pi U_1)$$

$$Z_2 \equiv \sqrt{-2 \log U_2} \sin(2\pi U_1)$$

so that $Z_1, Z_2 \overset{iid}{\sim} N(0, 1)$

What is this?

- **Inside the square root:** $-2 \log U_2 \sim \textit{Expo}\left(\frac{1}{2}\right) \sim \textit{Gamma}\left(\frac{2}{2}, \frac{1}{2}\right) \sim \chi^2_2$
- **Inside the cosine and sine:** $2\pi U_1$ is a uniformly distributed angle in polar coordinates, and the cos and sin convert this to the cartesian x and y components, respectively.
- **Altogether in polar coordinates:** The square root part is like a radius distributed $\chi \sim |Z|$, and the second part is an angle.

Key formulas we reviewed

CDF: $F_X(x) = P(X < x)$

PDF: $f_X(x) = \frac{\partial}{\partial x} F_X(x)$ (this $\neq P(X = x)$, which is 0!)

Definition of expectation: $E(X) = \begin{cases} \sum_x xp(x), & \text{if } x \text{ discrete} \\ \int xf(x)dx, & \text{if } x \text{ continuous} \end{cases}$

LOTUS (Law of the Unconscious Statistician):

$$E(g[X]) = \begin{cases} \sum_x g(x)p(x), & \text{if } x \text{ discrete} \\ \int g(x)f(x)dx, & \text{if } x \text{ continuous} \end{cases}$$

Adam's law (iterated expectation): $E(Y) = E(E[Y \mid X])$

Evve's law (total variance): $V(Y) = E(V[Y \mid X]) + V(E[Y \mid X])$

Key formulas we reviewed

Variance of sums:

$$V(aX + bY) = a^2X + b^2Y + 2abCov(X, Y)$$

$$V(aX - bY) = a^2x + b^2Y - 2abCov(X, Y)$$

Definition of variance:

$$\begin{aligned} V(X) &= E\left[(X - E[X])^2\right] \\ &= E\left(X^2 - [E(X)]^2\right) \end{aligned}$$

Definition of covariance:

$$\begin{aligned} Cov(X, Y) &= E\left[(X - E[X])\,(Y - E[Y])\right] \\ &= E\left(XY - E[X]E[Y]\right) \end{aligned}$$

Definition of correlation: $Cor(X, Y) = \frac{Cov(X,Y)}{SD(X)SD(Y)}$

Key resource: William Chen's Probability cheat sheet

http://www.wzchen.com/probability-cheatsheet/