Lecture 3

Ciprian Crainiceanu

Department of Biostatistics
Johns Hopkins Bloomberg School of Public Health
Johns Hopkins University

August 11, 2016

Table of contents

Outline

1. Define expected values
2. Properties of expected values
3. Unbiasedness of the sample mean
4. Define variances
5. Define the standard deviation
6. Calculate Bernoulli variance

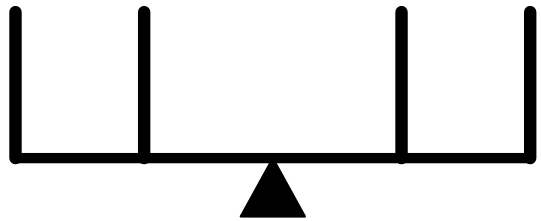
Expected values

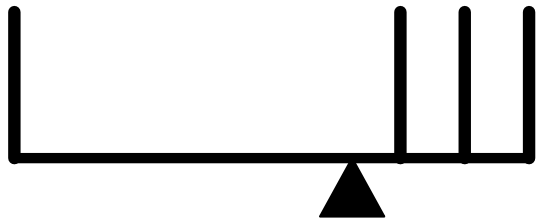
- The **expected value** or **mean** of a random variable is the center of mass of its distribution
- For discrete random variable X with PMF $p(x)$, it is defined as follows

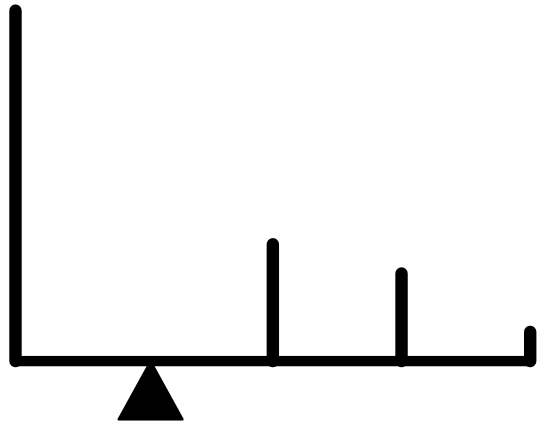
$$E[X] = \sum_x xp(x) = \sum_{\text{possible values of X}} xP(X = x)$$

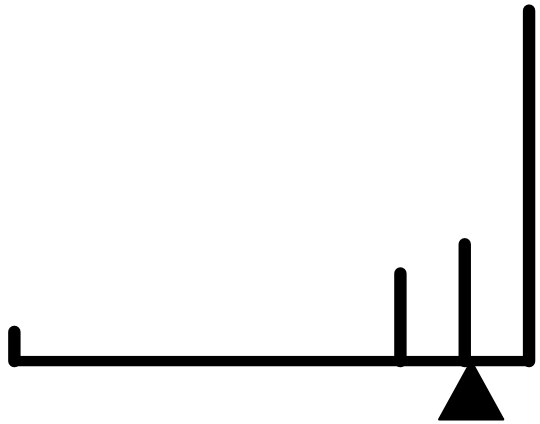
  where the sum is taken over the possible values of x
- $E[X]$ represents the center of mass of a collection of locations and weights, $\{x, p(x)\}$
- $E[X]$ is not necessarily among the values that the variable X takes

Example









Expected values

$$E[X] = \sum_x xp(x)$$

- $E(X_1) = (-4) \times 0.25 + (-3) \times 0.25 + 3 \times 0.25 + 4 \times 0.25 = 0$
- $E(X_2) = (-4) \times 0.25 + 1 \times 0.25 + 2 \times 0.25 + 3 \times 0.25 = 0.5$
- $E(X_3) = (-4) \times 0.60 + 1 \times 0.20 + 2 \times 0.15 + 3 \times 0.05 = -1.75$
- $E(X_4) = (-4) \times 0.05 + 1 \times 0.10 + 2 \times 0.15 + 3 \times 0.60 = 2$

Note that $\min(X) \leq E[X] \leq \max(X)$

Example: Fair coin

- Suppose a coin is flipped and X is declared 0 or 1 corresponding to a head or a tail, respectively
- What is the expected value of X?

$$E[X] = .5 \times 0 + .5 \times 1 = .5$$

- Note, if thought about geometrically, this answer is obvious; if two equal weights are spaced at 0 and 1, the center of mass will be .5

Example: Bernoulli

- Suppose that a person is infected with the flu virus with probability θ
- 1 is "person is infected", 0 is "person is not infected"
- X takes the value 1 with probability θ and 0 with probability $1 - \theta$
- What is the expected value of X?

$$E[X] = (1 - \theta) \times 0 + \theta \times 1 = \theta$$

- Interpretation: If one selects at random one person from a population of individuals with $100\theta\%$ infected people then the probability of this person to be infected is θ

Example: die rolls

- Suppose that a die is tossed and X is the number face up
- What is the expected value of X?
- X takes values in $\{1, 2, 3, 4, 5, 6\}$
- $P(X = 1) = \ldots = P(X = 6) = \frac{1}{6}$

$$E[X] = 1 \times \frac{1}{6} + 2 \times \frac{1}{6} + 3 \times \frac{1}{6} + 4 \times \frac{1}{6} + 5 \times \frac{1}{6} + 6 \times \frac{1}{6} = 3.5$$

- Again, the geometric argument makes this answer obvious without calculation
- Note that the mean **is not** the most likely value the variable takes. In this example $P\{X = E[X]\} = 0$

Example: die rolls

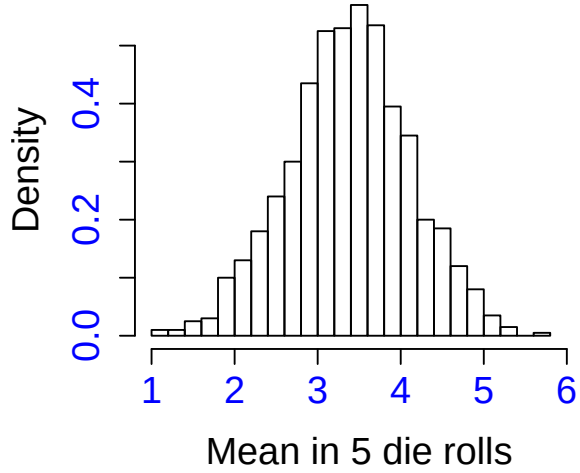
- Suppose that a die is tossed and X is the number face up
- What is the expected value of X^2?
- What is the expected value of $\sqrt{X}$?

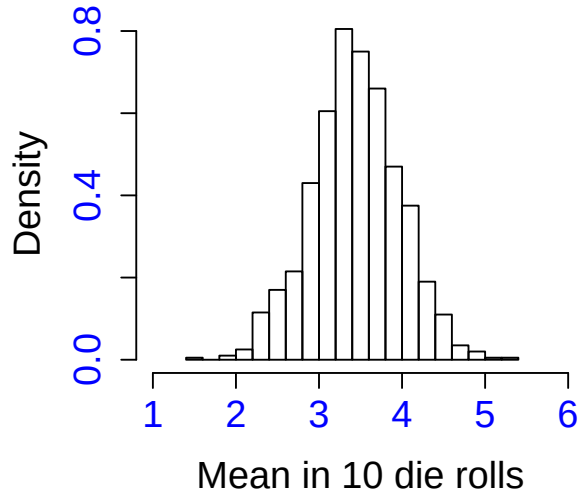
Using R to calculate the sample mean

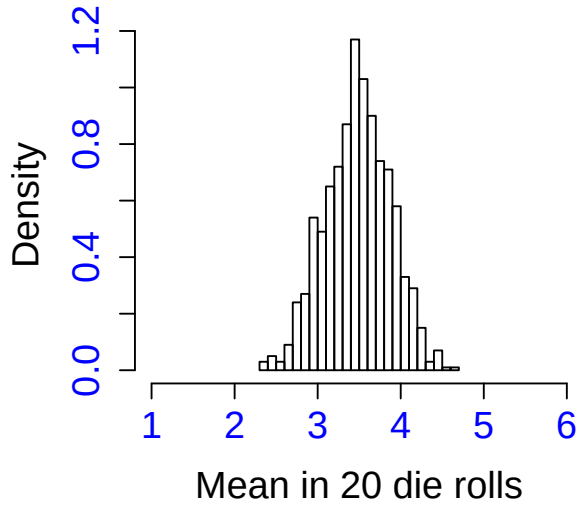
Simulate the mean of n die rolls

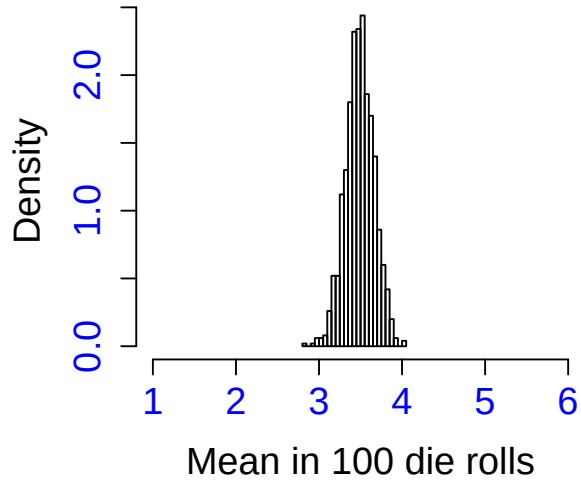
```
mx5 <- rep ( 0, 1000 )
mx10=mx5
mx20=mx5
mx100=mx5


for ( i in 1:1000 )
   {mx5[i] <- mean(sample(1:6,5,replace=T))
   mx10[i] <- mean(sample(1:6,10,replace=T))
   mx20[i] <- mean(sample(1:6,20,replace=T))
   mx100[i] <- mean(sample(1:6,100,replace=T))}
```

Example


Density
0.0
0.2
0.4
1
2
3
4
5
6
Mean in 5 die rolls


Density
0.0
0.4
0.8
1
2
3
4
5
6
Mean in 10 die rolls


Density
0.0
0.4
0.8
1.2
1
2
3
4
5
6
Mean in 20 die rolls


Density
0.0
1.0
2.0
1
2
3
4
5
6
Mean in 100 die rolls

Example: mental health

- Suppose that one observes the mental health of people in a large ($n = 10,000$) cohort study
- For each subject one observes one of the outcomes "healthy", "mild", "moderate", "serious"

What is the random variable and what is its mean?

Continuous random variables

- For a continuous random variable, X, with density, $f(x)$, the expected value is defined as follows

$$E[X] = \int_{-\infty}^{\infty} xf(x)dx$$

- This definition borrows from the definition of center of mass for a continuous body

Example: uniform distribution

- Consider that the random variable X has a distribution with density $f(x) = 1$ for $x \in [0, 1]$ and 0 otherwise
- Plot $f(x)$
- This is called the uniform distribution and we write $X \sim U[0, 1]$
- Is this a valid density?
- Suppose that X follows this density; what is its expected value?

$$E[X] = \int_0^1 x dx = \frac{x^2}{2} \Big|_0^1 = 1/2$$

- What is the pdf of the Uniform distribution on $[a, b]$?
- What is its mean?

Example: uniform distribution

```
y1<-mean(runif(10000))
y2<-mean(runif(10000,0,10))
y3<-mean(runif(10000,-2,15))
```

- If $X \sim U(a, b)$ calculate $E[X]$
- If $X_1, \ldots, X_n$ are independent random variables with pdf $f(x)$

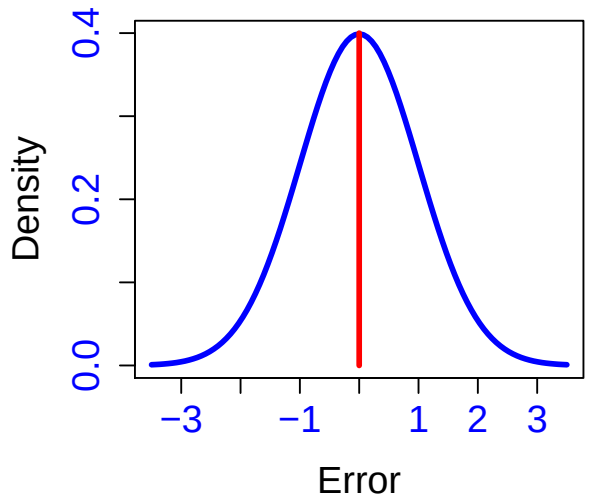
$$\overline{X}_n = \frac{1}{n}(X_1 + \ldots X_n) \approx E[X] = \int x f(x) dx$$

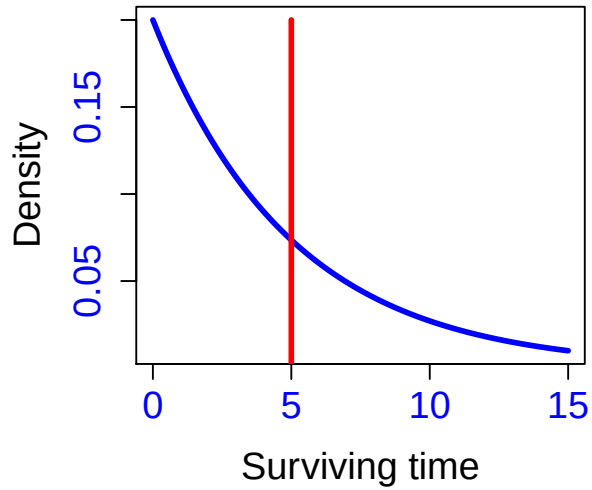
- When n is larger the approximation is better (aka the strong law of large numbers)

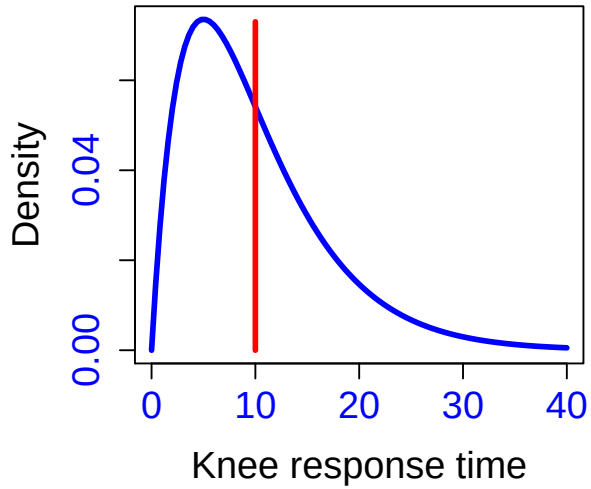
Distributions: the mean

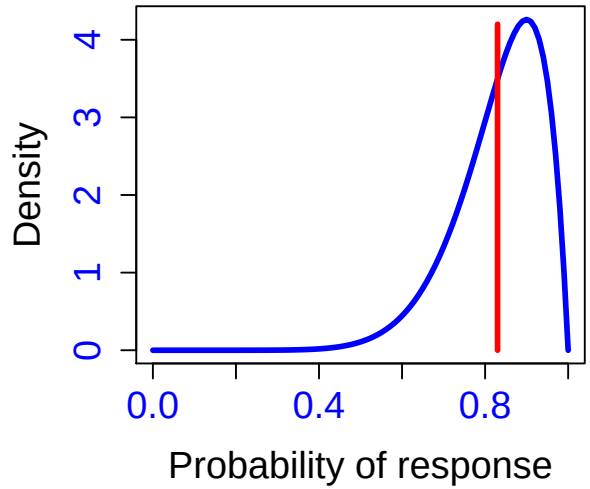
- The mean is the "center of mass"; may not even be an acceptable value for experiment
- The mean is not typically the median
- For symmetric distributions (aka error distributions) the mean and the median are equal
- The mean can be heavily influenced by skewness, outliers
- Examples: medical expenditure in the US, average net worth, insurance claims
- **Remember**: interpretation, interpretation, interpretation

Example


Density
0.0
0.2
0.4
−3
−1
1
2
3
Error


Density
0.05
0.15
0
5
10
15
Surviving time


Density
0.00
0.04
0
10
20
30
40
Knee response time


Density
0
1
2
3
4
0.0
0.4
0.8
Probability of response

Example: Gamma distribution

- Consider that the random variable X has a distribution with density

$$f(x) = \frac{1}{\beta^{\alpha}\Gamma(\alpha)} x^{\alpha-1} e^{-x/\beta}, \text{ for every } x \in (0, \infty)$$

- $\Gamma(\alpha) = \int_0^{\infty} x^{\alpha-1} e^{-x} dx$
- Is this a valid density?
- Suppose that X follows this density; what is $E[X]$?
- What is $E[X^{\pi}]$?

Rules about expected values

- The expected value is a linear operator
- If a and b are not random and X and Y are two random variables then
  - $E[aX + b] = aE[X] + b$
  - $E[X + Y] = E[X] + E[Y]$
  - $X \sim exp(5)$, $Y \sim U(20, 25)$. $E[3X + 2Y] =?$
- *In general* if g is a function that is not linear,

$$E[g(X)] \neq g(E[X])$$

- For example, in general, $E[X^2] \neq E[X]^2$, $E[\log(X)] \neq \log(E[X])$

Nonlinear transformation of the mean

$Var = E(X^2) - E^2(X) \geq 0$

- We show that $E[X]^2 \leq E[X^2]$ for every discrete random variable X
- This is actually true for every random variable
- We need to show $\{\sum_x xp(x)\}^2 \leq \sum_x x^2 p(x)$
- It is known that $(\sum_i a_i b_i)^2 \leq (\sum_i a_i^2)(\sum_i b_i^2)$
- Take $a_x = x\sqrt{p(x)}$, $b_x = \sqrt{p(x)}$
- Show that $(a_1 b_1 + a_2 b_2)^2 \leq (a_1^2 + a_2^2)(b_1^2 + b_2^2)$

In general, if $h(\cdot)$ is a convex function $h(E[X]) \leq E[h(X)]$

Example

- You flip a coin, X and simulate a uniform random number Y, what is the expected value of their sum?

$$E[X + Y] = E[X] + E[Y] = .5 + .5 = 1$$

- Another example, you roll a die twice. What is the expected value of the average?
- Let X_1 and X_2 be the results of the two rolls

$$E[(X_1 + X_2)/2] = \frac{1}{2}(E[X_1] + E[X_2]) = \frac{1}{2}(3.5 + 3.5) = 3.5$$

Example

X_1	X_2	$P(X)$	$\frac{X_1+X_2}{2}$
1	1	1/36	1
1	2	1/36	1.5
1	3	1/36	2
1	4	1/36	2.5
1	5	1/36	3
1	6	1/36	3.5
2	1	1/36	1.5
2	2	1/36	2
2	3	1/36	2.5
2	4	1/36	3
⋮	⋮	⋮	⋮

- What is the probability of getting $\frac{X_1+X_2}{2} = 3.5$?
- How many values does $\frac{X_1+X_2}{2}$ take?

Example

$\frac{X_1+X_2}{2}$	1	1.5	2	2.5	3	3.5	4	...
$P\left(\frac{X_1+X_2}{2}=x\right)$	$\frac{1}{36}$	$\frac{2}{36}$	$\frac{3}{36}$	$\frac{4}{36}$	$\frac{5}{36}$	$\frac{6}{36}$	$\frac{5}{36}$	...

$$E[X] = 1 \times \frac{1}{36} + 1.5 \times \frac{2}{36} + 2 \times \frac{3}{36} + \dots$$

How many values does $\overline{X}_n$ take?

- min is 1, max is 6
- all values in increments of $1/n$

```
length(seq(1,6,by=1/20))
length(seq(1,6,by=1/200))
```

Example

1. Let X_i for $i = 1, \ldots, n$ be a collection of random variables, each from a distribution with mean μ
2. Calculate the expected value of the sample average of the X_i

$$\begin{aligned} E\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] &= \frac{1}{n} E\left[\sum_{i=1}^{n} X_i\right] \\ &= \frac{1}{n}\sum_{i=1}^{n} E\left[X_i\right] \\ &= \frac{1}{n}\sum_{i=1}^{n} \mu = \mu. \end{aligned}$$

Remark

- Therefore, the expected value of the **sample mean** is the **population mean** that it's trying to estimate
- When the expected value of an estimator is what it is trying to estimate, we say that the estimator is **unbiased**
- An estimator is any function of the data, $U(X)$. A parameter is any unknown quantity in the model, θ
- $U(X)$ is unbiased for θ if $E[U(X)] = \theta$

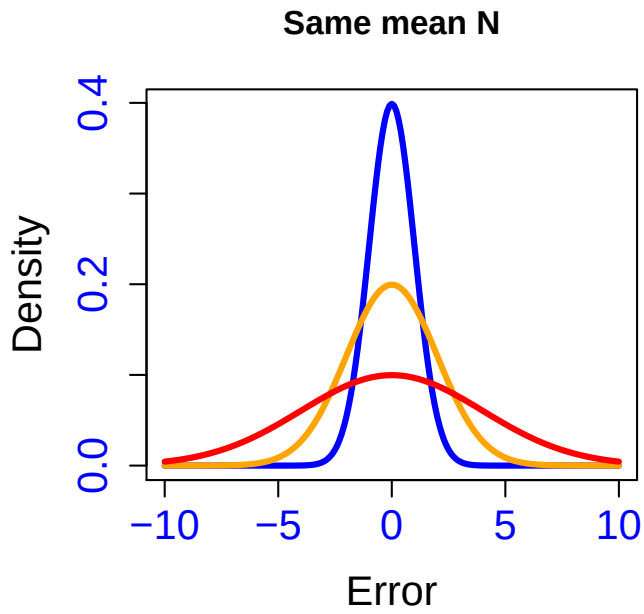
The variance

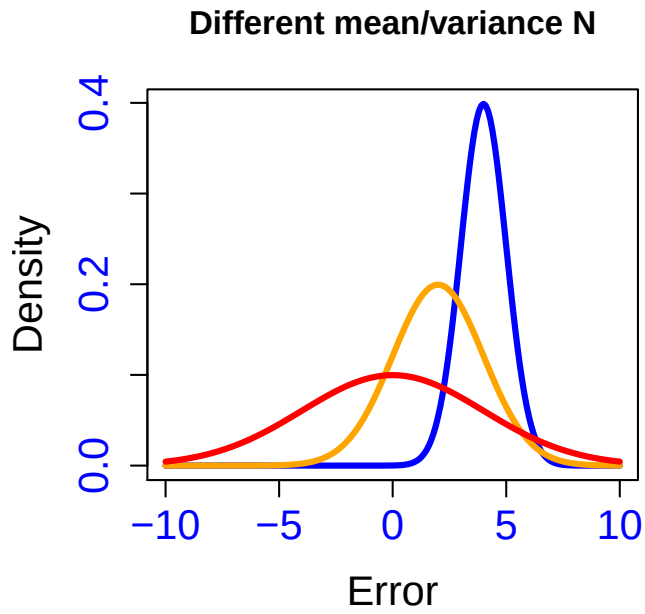
- The variance of a random variable is a measure of *spread*
- If X is a random variable with mean μ, the variance of X is defined as

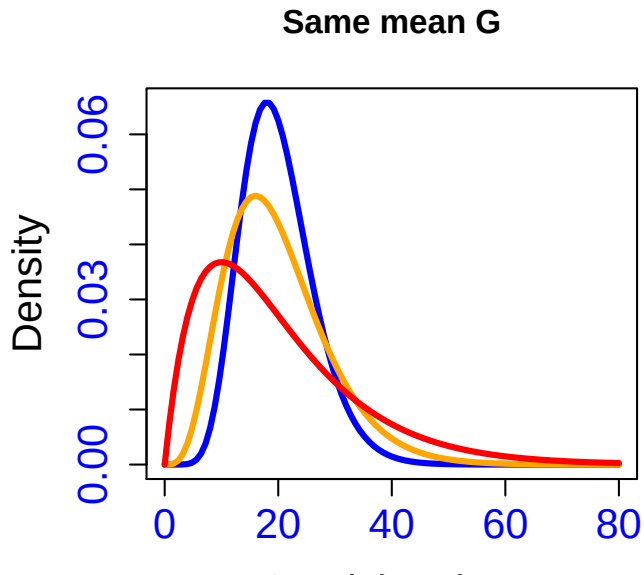
$$\mathrm{Var}(X) = E[(X - \mu)^2]$$

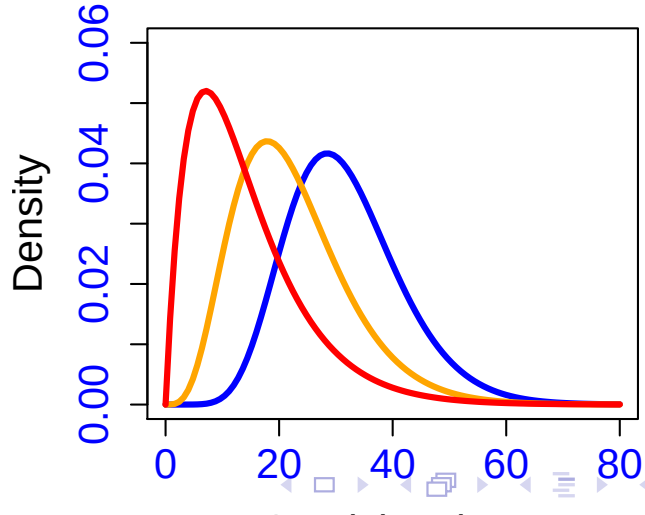
  the expected (squared) distance from the mean
- If X is a discrete random variable $\mathrm{Var}[X] = \sum_x (x - \mu)^2 p(x)$
- If X is a continuous random variable $\mathrm{Var}[X] = \int_{-\infty}^{\infty} (x - \mu)^2 f(x) dx$
- Densities with a higher variance are more spread out than densities with a lower variance

Example


Same mean N
Density
0.0
0.2
0.4
−10
−5
0
5
10
Error


Different mean/variance N
Density
0.0
0.2
0.4
−10
−5
0
5
10
Error


Same mean G
Density
0.00
0.03
0.06
0
20
40
60
80


Same variance G
Density
0.00
0.02
0.04
0.06
0
20
40
60
80

The variance

- Convenient computational form

$$\mathrm{Var}(X) = E[X^2] - E[X]^2$$

- If a is constant then $\mathrm{Var}(aX) = a^2\mathrm{Var}(X)$
- The square root of the variance is called the **standard deviation**
- $\mathrm{SD}(X) = \sqrt{\mathrm{Var}(X)}$
- The standard deviation has the same units as X
- If a is constant then $\mathrm{SD}(aX) = a\mathrm{SD}(X)$: scale invariant
- Prove these results

Variance: some R code

```
x=seq(-10,10,length=201)
y1<-dnorm(x)
y2<-dnorm(x,0,2)
y3<-dnorm(x,0,4)
plot(x,y1,type="l",col="blue",lwd=3)
lines(x,y2,col="orange",lwd=3)
lines(x,y3,col="red",lwd=3)
```

Calculating the empirical variance in R

```
y<-rnorm(100,0,4)
var(y)
sd(y)
```

Example

- What is the sample variance from the result of a toss of a die?
- $\mathrm{Var}(X) = E[X^2] - E[X]^2$
- What are the values that X^2 can take?

$Y = X^2$	1	4	9	16	25	36
$P(Y = y)$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$

Example

- What is the sample variance from the result of a toss of a die?
  - $E[X] = 3.5$
  - $E[X^2] = 1^2 \times \frac{1}{6} + 2^2 \times \frac{1}{6} + 3^2 \times \frac{1}{6} + 4^2 \times \frac{1}{6} + 5^2 \times \frac{1}{6} + 6^2 \times \frac{1}{6} = 15.17$
- $\mathrm{Var}(X) = E[X^2] - E[X]^2 \approx 2.92$

```
x=1:6
ex2=sum(x^2*rep(1/6,6))
ex=sum(x*rep(1/6,6))
varx=ex2-ex^2
```

Example

- What's the sample variance from the result of the toss of a coin with probability of heads (1) of p?
  - $E[X] = 0 \times (1-p) + 1 \times p = p$
  - $E[X^2] = E[X] = p$
- $\mathrm{Var}(X) = E[X^2] - E[X]^2 = p - p^2 = p(1-p)$

Example

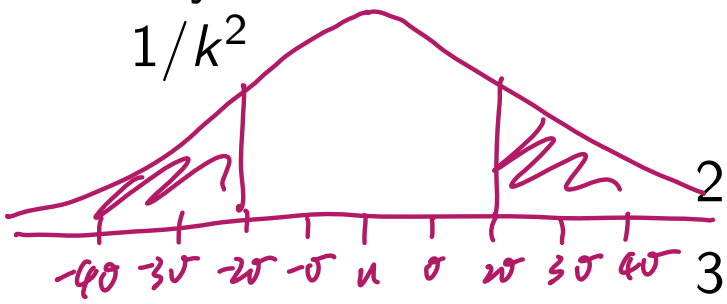
- Suppose that a random variable is such that $0 \leq X \leq 1$ and $E[X] = p$
- Note that $X^2 \leq X$ so that $E[X^2] \leq E[X] = p$
- $\mathrm{Var}(X) = E[X^2] - E[X]^2 \leq E[X] - E[X]^2 = p(1-p)$
- Therefore the Bernoulli variance is the largest possible for random variables bounded between 0 and 1
- Largest variance is attained at $p = 0.5$
- For every Beta distribution there exists a Bernoulli distribution with the same mean and larger variance

Interpreting variances

- Chebyshev's inequality is useful for interpreting variances
- This inequality states that

$$P(|X - \mu| \geq k\sigma) \leq \frac{1}{k^2}$$

- For example, the probability that a random variable lies beyond k standard deviations from its mean is less than $1/k^2$



$$2\sigma \rightarrow 25\%$$
$$3\sigma \rightarrow 11\%$$
$$4\sigma \rightarrow 6\%$$

- Note that this is only an upper bound; the actual probability might be quite a bit smaller

Proof of Chebyshev's inequality

$$
\begin{aligned}
P(|X-\mu| > k\sigma) &= \int_{\{x:|x-\mu|>k\sigma\}} f(x)dx \\
&\leq \int_{\{x:|x-\mu|>k\sigma\}} \frac{(x-\mu)^2}{k^2\sigma^2} f(x)dx \\
&\leq \int_{-\infty}^{\infty} \frac{(x-\mu)^2}{k^2\sigma^2} f(x)dx \\
&= \frac{1}{k^2}
\end{aligned}
$$

Comparing Chebyshev with parametric assumptions

$$P(|X - \mu| \geq k\sigma) \leq \frac{1}{k^2}$$

$k\sigma$	2	3	4	5
Any	0.250	0.111	0.063	0.040
t(3)	0.041	0.014	0.006	0.003
Gamma(2,2)	0.046	0.014	0.004	0.001
Normal	0.046	0.003	6.33×10^{-5}	5.73×10^{-7}

Comparing Chebyshev with parametric assumptions

```
k=2:5 # Multiples of SD

2*(1-pnorm(k)) # Normal

sdt3=sqrt(3) # SD of a t(3) distribution
2*(1-pt(k*sdt3,df=3)) # t(3)

sh=2 # shape of Gamma(sh,sc)
sc=2 # scale of Gamma(sh,sc)
m=sh*sc # mean of Gamma(sh,sc)
sdg = sqrt(sh*sc^2) # SD of Gamma(sh,sc)
pgamma(m-k*sdg,shape=sh,scale=sc)+
1-pgamma(m+k*sdg,shape=sh,scale=sc)
```

Example

- IQs are often said to be distributed with a mean of 100 and a sd of 15
- What is the probability of a randomly drawn person having an IQ higher than 160 or below 40?
- Thus we want to know the probability of a person being more than 4 standard deviations from the mean
- Thus Chebyshev's inequality suggests that this will be no larger than 6%
- IQs distributions are often cited as being bell shaped, in which case this bound is very conservative
- The probability of a random draw from a bell curve being 4 standard deviations from the mean is on the order of 10^{-5} (one thousandth of one percent)

Example

- A popular buzz phrase in industrial quality control is Motorola's"Six Sigma" whereby businesses are suggested to control extreme events or rare defective parts
- Chebyshev's inequality states that the probability of a "Six Sigma" event is less than $1/6^2 \approx 3\%$
- If a bell curve is assumed, the probability of a "six sigma" event is on the oder of 10^{-9} (one ten millionth of a percent)

Coefficient of variation

- For a variable $X > 0$

$$C_v(X) = \frac{\text{SD}(X)}{E[X]}$$

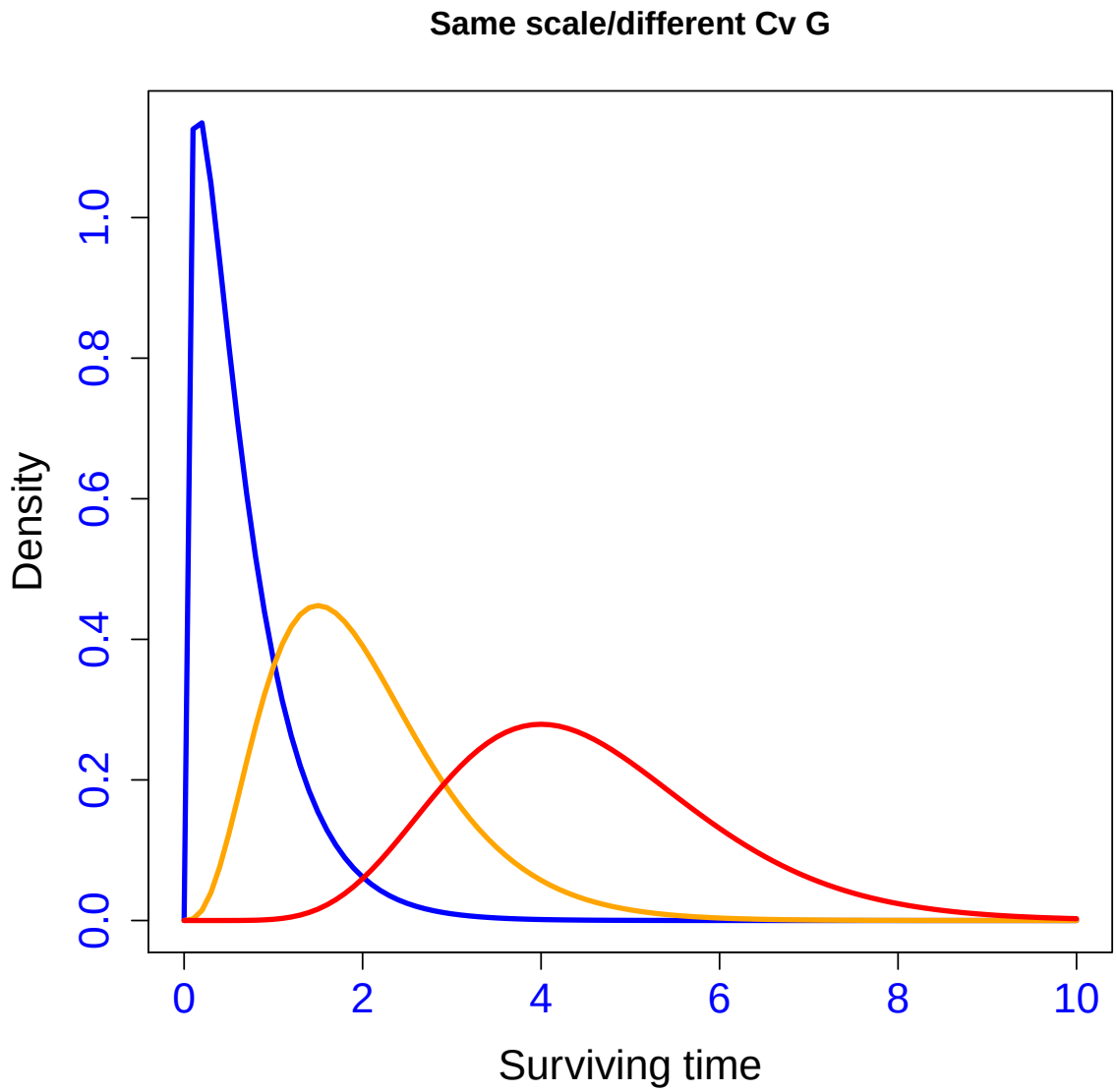
- Measures the amount of variability (as described by the standard deviation) relative to the mean
- C_v is the inverse of the signal-to-noise ratio
- It does not have units (mean and standard deviation are on the same scale)
- Strong link to Cohen's d used in sample size calculations

$$C_d = \frac{E[X_1] - E[X_2]}{\text{SD}}$$

Coefficient of variation

```
x=seq(0,10,length=101) # set the grid
# Gamma with C_v = 1/sqrt(1.3) = 0.88
y1<-dgamma(x,shape=1.3,scale=1/2)
# Gamma with C_v = 1/sqrt(4) = 0.50
y2<-dgamma(x,shape=4,scale=1/2)
# Gamma with C_v = 1/sqrt(9) = 0.33
y3<-dgamma(x,shape=9,scale=1/2)

plot(x,y1,type="l",col="blue",lwd=3)
lines(x,y2,col="orange",lwd=3)
lines(x,y3,col="red",lwd=3)
```

Varying C_v


Same scale/different Cv G
Density
0.0
0.2
0.4
0.6
0.8
1.0
0
2
4
6
8
10
Surviving time