Introduction to statistical inference

Statistical inference

Brian Caffo, Jeff Leek, Roger Peng
Johns Hopkins Bloomberg School of Public Health

Statistical inference defined

Statistical inference is the process of drawing formal conclusions from data.

In our class, we wil define formal statistical inference as settings where one wants to infer facts about a population using noisy statistical data where uncertainty must be accounted for.

Motivating example: who's going to win the election?

In every major election, pollsters would like to know, ahead of the actual election, who's going to win. Here, the target of estimation (the estimand) is clear, the percentage of people in a particular group (city, state, county, country or other electoral grouping) who will vote for each candidate.

We can not poll everyone. Even if we could, some polled may change their vote by the time the election occurs. How do we collect a reasonable subset of data and quantify the uncertainty in the process to produce a good guess at who will win?

Motivating example: is hormone replacement therapy effective?

A large clinical trial (the Women's Health Initiative) published results in 2002 that contradicted prior evidence on the efficacy of hormone replacement therapy for post menopausal women and suggested a negative impact of HRT for several key health outcomes. **Based on a statistically based protocol, the study was stopped early due an excess number of negative events.**

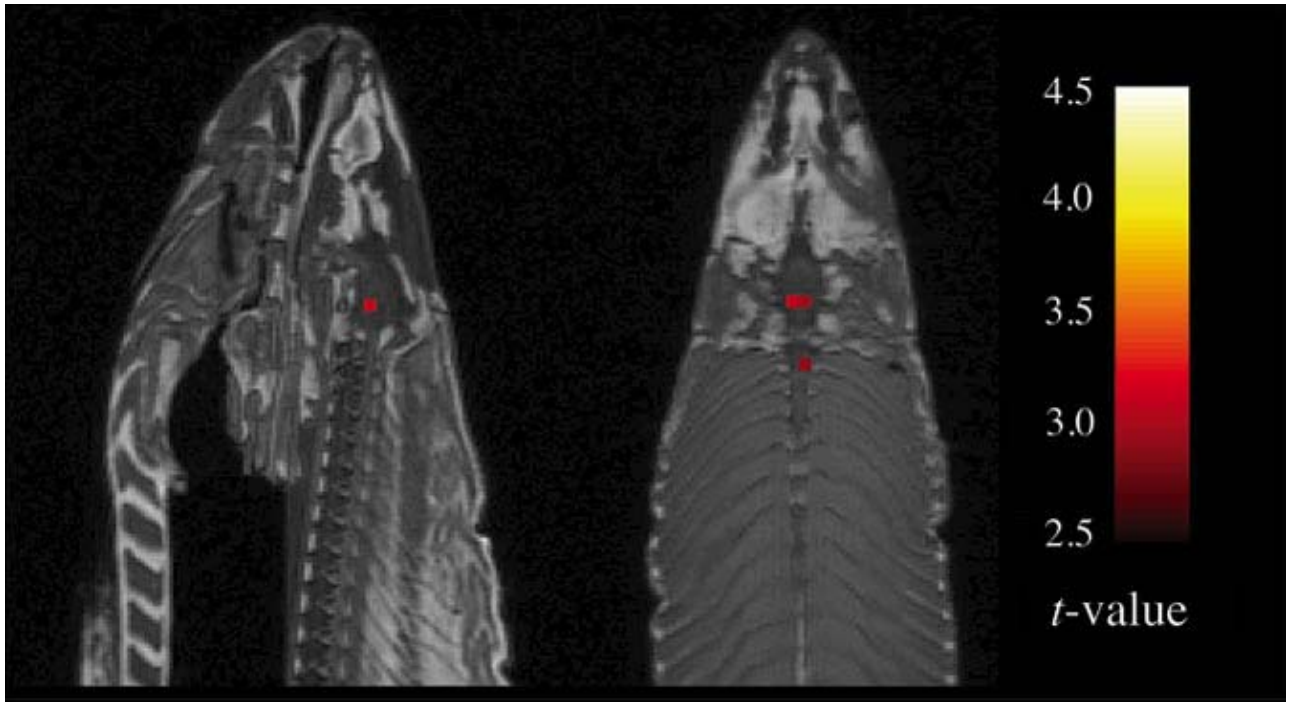
Here's there's two inferential problems.

1. Is HRT effective?
2. How long should we continue the trial in the presence of contrary evidence?

See WHI writing group paper JAMA 2002, Vol 288:321 - 333. for the paper and Steinkellner et al. Menopause 2012, Vol 19:616 621 for adiscussion of the long term impacts

Motivating example

Brain activation



http://www.wired.com/2009/09/fmrisalmon/

Summary

- These examples illustrate many of the difficulties of trying to use data to create general conclusions about a population.
- Paramount among our concerns are:
  - Is the sample representative of the population that we'd like to draw inferences about?
  - Are there known and observed, known and unobserved or unknown and unobserved variables that contaminate our conclusions?
  - Is there systematic bias created by missing data or the design or conduct of the study?
  - What randomness exists in the data and how do we use or adjust for it? Here randomness can either be explicit via randomization or random sampling, or implicit as the aggregation of many complex uknown processes.
  - Are we trying to estimate an underlying mechanistic model of phenomena under study?
- Statistical inference requires navigating the set of assumptions and tools and subsequently thinking about how to draw conclusions from data.

Example goals of inference

1. Estimate and quantify the uncertainty of an estimate of a population quantity (the proportion of people who will vote for a candidate).
2. Determine whether a population quantity is a benchmark value ("is the treatment effective?").
3. Infer a mechanistic relationship when quantities are measured with noise ("What is the slope for Hooke's law?")
4. Determine the impact of a policy? ("If we reduce polution levels, will asthma rates decline?")
5. Talk about the probability that something occurs.

Example tools of the trade

1. Randomization: concerned with balancing unobserved variables that may confound inferences of interest
2. Random sampling: concerned with obtaining data that is representative of the population of interest
3. Sampling models: concerned with creating a model for the sampling process, the most common is so called "iid".
4. Hypothesis testing: concerned with decision making in the presence of uncertainty
5. Confidence intervals: concerned with quantifying uncertainty in estimation
6. Probability models: a formal connection between the data and a population of interest. Often probability models are assumed or are approximated.
7. Study design: the process of designing an experiment to minimize biases and variability.
8. Nonparametric bootstrapping: the process of using the data to, with minimal probability model assumptions, create inferences.
9. Permutation, randomization and exchangeability testing: the process of using data permutations to perform inferences.

Different thinking about probability leads to different styles of inference

We won't spend too much time talking about this, but there are several different styles of inference. Two broad categories that get discussed a lot are:

1. Frequency probability: is the long run proportion of times an event occurs in independent, identically distributed repetitions.
2. Frequency inference: uses frequency interpretations of probabilities to control error rates. Answers questions like "What should I decide given my data controlling the long run proportion of mistakes I make at a tolerable level."
3. Bayesian probability: is the probability calculus of beliefs, given that beliefs follow certain rules.
4. Bayesian inference: the use of Bayesian probability representation of beliefs to perform inference. Answers questions like "Given my subjective beliefs and the objective information from the data, what should I believe now?"

Data scientists tend to fall within shades of gray of these and various other schools of inference.

In this class

- In this class, we will primarily focus on basic sampling models, basic probability models and frequency style analyses to create standard inferences.
- Being data scientists, we will also consider some inferential strategies that rely heavily on the observed data, such as permutation testing and bootstrapping.
- As probability modeling will be our starting point, we first build up basic probability.

Where to learn more on the topics not covered

1. Explicit use of random sampling in inferences: look in references on "finite population statistics". Used heavily in polling and sample surveys.
2. Explicit use of randomization in inferences: look in references on "causal inference" especially in clinical trials.
3. Bayesian probability and Bayesian statistics: look for basic itroductory books (there are many).
4. Missing data: well covered in biostatistics and econometric references; look for references to "multiple imputation", a popular tool for addressing missing data.
5. Study design: consider looking in the subject matter area that you are interested in; some examples with rich histories in design:
    1. The epidemiological literature is very focused on using study design to investigate public health.
    2. The classical development of study design in agriculture broadly covers design and design principles.
    3. The industrial quality control literature covers design thoroughly.

Probability

Statistical Inference

Brian Caffo, Jeff Leek, Roger Peng
Johns Hopkins Bloomberg School of Public Health

Probability

- In these slides we will cover the basics of probability at low enough level to have a basic understanding for the rest of the series
- For a more complete treatment see the class Mathematical Biostatistics Boot Camp 1
  - Youtube: www.youtube.com/playlist?list=PLpl-gQkQivXhk6qSyiNj51qamjAtZISJ-
  - Coursera: www.coursera.org/course/biostats
  - Git: http://github.com/bcaffo/Caffo-Coursera

Probability

Given a random experiment (say rolling a die) a probability measure is a population quantity that summarizes the randomness.

Specifically, probability takes a possible outcome from the expertiment and:

- assigns it a number between 0 and 1
- so that the probability that something occurs is 1 (the die must be rolled) and
- so that the probability of the union of any two sets of outcomes that have nothing in common (mutually exclusive) is the sum of their respective probabilities.

The Russian mathematician Kolmogorov formalized these rules.

Rules probability must follow

- The probability that nothing occurs is 0
- The probability that something occurs is 1
- The probability of something is 1 minus the probability that the opposite occurs
- The probability of at least one of two (or more) things that can not simultaneously occur (mutually exclusive) is the sum of their respective probabilities
- If an event A implies the occurrence of event B, then the probability of A occurring is less than the probability that B occurs
- For any two events the probability that at least one occurs is the sum of their probabilities minus their intersection.

Example

The National Sleep Foundation (www.sleepfoundation.org) reports that around 3% of the American population has sleep apnea. They also report that around 10% of the North American and European population has restless leg syndrome. Does this imply that 13% of people will have at least one sleep problems of these sorts?

Example continued

Answer: No, the events can simultaneously occur and so are not mutually exclusive. To elaborate let:

If you want to see the mathematics

$$A_1 = \{\text{Person has sleep apnea}\}$$
$$A_2 = \{\text{Person has RLS}\}$$

Then

$$\begin{aligned} P(A_1 \cup A_2) &= P(A_1) + P(A_2) - P(A_1 \cap A_2) \\ &= 0.13 - \text{Probability of having both} \end{aligned}$$

Likely, some fraction of the population has both.

Going further

Probability calculus is useful for understanding the rules that probabilities must follow.

However, we need ways to model and think about probabilities for numeric outcomes of experiments (broadly defined).

Densities and mass functions for random variables are the best starting point for this.

Remember, everything we're talking about up to at this point is a population quantity not a statement about what occurs in the data.

- We're going with this is: use the data to estimate properties of the population.

Random variables

- A **random variable** is a numerical outcome of an experiment.
- The random variables that we study will come in two varieties, **discrete** or **continuous**.
- Discrete random variable are random variables that take on only a countable number of possibilities and we talk about the probability that they take specific values
- Continuous random variable can conceptually take any value on the real line or some subset of the real line and we talk about the probability that they line within some range

Examples of variables that can be thought of as random variables

Experiments that we use for intuition and building context

- The $(0-1)$ outcome of the flip of a coin
- The outcome from the roll of a die

Specific instances of treating variables as if random

- The web site traffic on a given day
- The BMI of a subject four years after a baseline measurement
- The hypertension status of a subject randomly drawn from a population
- The number of people who click on an ad
- Intelligence quotients for a sample of children

PMF

A probability mass function evaluated at a value corresponds to the probability that a random variable takes that value. To be a valid pmf a function, p, must satisfy

1. It must always be larger than or equal to 0.
2. The sum of the possible values that the random variable can take has to add up to one.

Example

Let X be the result of a coin flip where $X = 0$ represents tails and $X = 1$ represents heads.

$$p(x) = (1/2)^x (1/2)^{1-x} \quad \text{for} \quad x = 0, 1$$

Suppose that we do not know whether or not the coin is fair; Let θ be the probability of a head expressed as a proportion (between 0 and 1).

$$p(x) = \theta^x (1-\theta)^{1-x} \quad \text{for} \quad x = 0, 1$$

PDF

A probability density function (pdf), is a function associated with a continuous random variable

Areas under pdfs correspond to probabilities for that random variable

To be a valid pdf, a function must satisfy

1. It must be larger than or equal to zero everywhere.
2. The total area under it must be one.

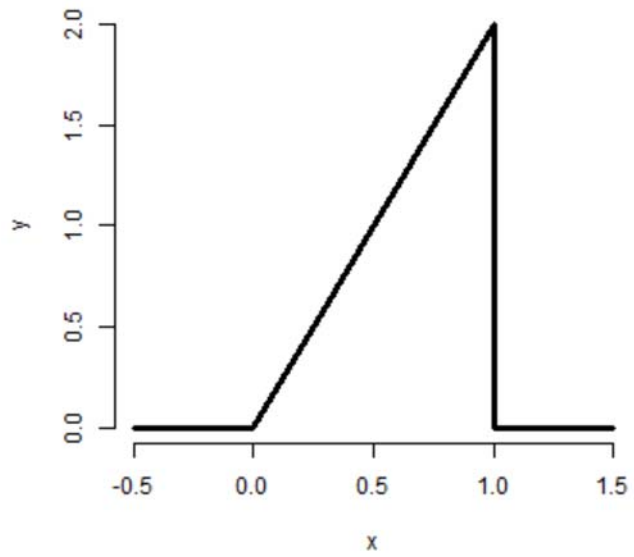
Example

Suppose that the proportion of help calls that get addressed in a random day by a help line is given by

$$f(x) = \begin{cases} 2x & \text{for} \quad 0 < x < 1 \\ 0 & \text{otherwise} \end{cases}$$

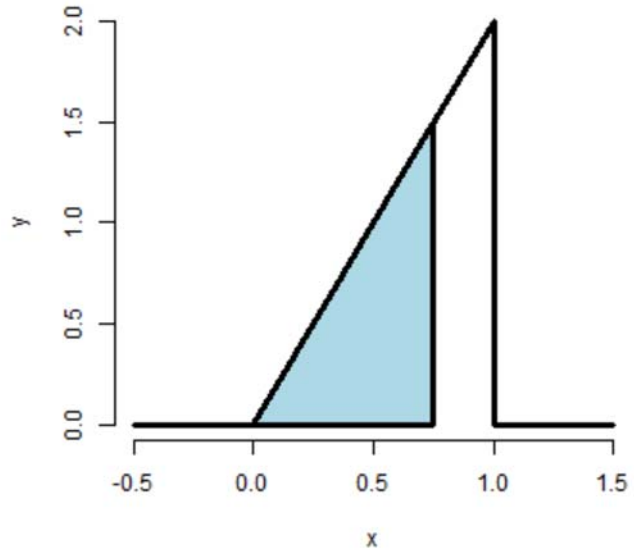
Is this a mathematically valid density?

```r
x <- c(-0.5, 0, 1, 1, 1.5)
y <- c(0, 0, 2, 0, 0)
plot(x, y, lwd = 3, frame = FALSE, type = "l")
```



Example continued

What is the probability that 75% or fewer of calls get addressed?



```
1.5 * 0.75/2
```

```
## [1] 0.5625
```

```
pbeta(0.75, 2, 1)
```

```
## [1] 0.5625
```

CDF and survival function

Certain areas are so useful, we give them names

- The **cumulative distribution function** (CDF) of a random variable, X, returns the probability that the random variable is less than or equal to the value x

$$F(x) = P(X \leq x)$$

  (This definition applies regardless of whether X is discrete or continuous.)

- The **survival function** of a random variable X is defined as the probability that the random variable is greater than the value x

$$S(x) = P(X > x)$$

- Notice that $S(x) = 1 - F(x)$

Example

What are the survival function and CDF from the density considered before?

For $1 \geq x \geq 0$

$$F(x) = P(X \leq x) = \frac{1}{2} \, Base \times Height = \frac{1}{2} \, (x) \times (2x) = x^2$$

$$S(x) = 1 - x^2$$

```
pbeta(c(0.4, 0.5, 0.6), 2, 1)
```

```
## [1] 0.16 0.25 0.36
```

Quantiles

You've heard of sample quantiles. If you were the 95th percentile on an exam, you know that 95% of people scored worse than you and 5% scored better. These are sample quantities. Here we define their population analogs.

Definition

The α^{th} **quantile** of a distribution with distribution function F is the point x_α so that

$$F(x_\alpha) = \alpha$$

- A **percentile** is simply a quantile with α expressed as a percent
- The **median** is the 50^{th} percentile

For example

The 95^{th} percentile of a distribution is the point so that:

- the probability that a random variable drawn from the population is less is 95%
- the probability that a random variable drawn from the population is more is 5%

Example

What is the median of the distribution that we were working with before?

- We want to solve $0.5 = F(x) = x^2$
- Resulting in the solution

```
sqrt(0.5)
```

```
## [1] 0.7071
```

- Therefore, about 0.7071 of calls being answered on a random day is the median.

Example continued

R can approximate quantiles for you for common distributions

```
qbeta(0.5, 2, 1)
```

```
## [1] 0.7071
```

Summary

- You might be wondering at this point "I've heard of a median before, it didn't require integration. Where's the data?"
- We're referring to are **population quantities**. Therefore, the median being discussed is the **population median**.
- A probability model connects the data to the population using assumptions.
- Therefore the median we're discussing is the **estimand**, the sample median will be the **estimator**

Probability

Statistical Inference

Brian Caffo, Jeff Leek, Roger Peng
Johns Hopkins Bloomberg School of Public Health

Probability

- In these slides we will cover the basics of probability at low enough level to have a basic understanding for the rest of the series
- For a more complete treatment see the class Mathematical Biostatistics Boot Camp 1
  - Youtube: www.youtube.com/playlist?list=PLpl-gQkQivXhk6qSyiNj51qamjAtZISJ-
  - Coursera: www.coursera.org/course/biostats
  - Git: http://github.com/bcaffo/Caffo-Coursera

Probability

Given a random experiment (say rolling a die) a probability measure is a population quantity that summarizes the randomness.

Specifically, probability takes a possible outcome from the expertiment and:

- assigns it a number between 0 and 1
- so that the probability that something occurs is 1 (the die must be rolled) and
- so that the probability of the union of any two sets of outcomes that have nothing in common (mutually exclusive) is the sum of their respective probabilities.

The Russian mathematician Kolmogorov formalized these rules.

Rules probability must follow

- The probability that nothing occurs is 0
- The probability that something occurs is 1
- The probability of something is 1 minus the probability that the opposite occurs
- The probability of at least one of two (or more) things that can not simultaneously occur (mutually exclusive) is the sum of their respective probabilities
- If an event A implies the occurrence of event B, then the probability of A occurring is less than the probability that B occurs
- For any two events the probability that at least one occurs is the sum of their probabilities minus their intersection.

Example

The National Sleep Foundation (www.sleepfoundation.org) reports that around 3% of the American population has sleep apnea. They also report that around 10% of the North American and European population has restless leg syndrome. Does this imply that 13% of people will have at least one sleep problems of these sorts?

Example continued

Answer: No, the events can simultaneously occur and so are not mutually exclusive. To elaborate let:

If you want to see the mathematics

$$A_1 = \{\text{Person has sleep apnea}\}$$
$$A_2 = \{\text{Person has RLS}\}$$

Then

$$\begin{aligned} P(A_1 \cup A_2) &= P(A_1) + P(A_2) - P(A_1 \cap A_2) \\ &= 0.13 - \text{Probability of having both} \end{aligned}$$

Likely, some fraction of the population has both.

Going further

Probability calculus is useful for understanding the rules that probabilities must follow.

However, we need ways to model and think about probabilities for numeric outcomes of experiments (broadly defined).

Densities and mass functions for random variables are the best starting point for this.

Remember, everything we're talking about up to at this point is a population quantity not a statement about what occurs in the data.

- We're going with this is: use the data to estimate properties of the population.

Random variables

- A **random variable** is a numerical outcome of an experiment.
- The random variables that we study will come in two varieties, **discrete** or **continuous**.
- Discrete random variable are random variables that take on only a countable number of possibilities and we talk about the probability that they take specific values
- Continuous random variable can conceptually take any value on the real line or some subset of the real line and we talk about the probability that they line within some range

Examples of variables that can be thought of as random variables

Experiments that we use for intuition and building context

- The $(0-1)$ outcome of the flip of a coin
- The outcome from the roll of a die

Specific instances of treating variables as if random

- The web site traffic on a given day
- The BMI of a subject four years after a baseline measurement
- The hypertension status of a subject randomly drawn from a population
- The number of people who click on an ad
- Intelligence quotients for a sample of children

PMF

A probability mass function evaluated at a value corresponds to the probability that a random variable takes that value. To be a valid pmf a function, p, must satisfy

1. It must always be larger than or equal to 0.
2. The sum of the possible values that the random variable can take has to add up to one.

Example

Let X be the result of a coin flip where $X = 0$ represents tails and $X = 1$ represents heads.

$$p(x) = (1/2)^x (1/2)^{1-x} \quad \text{for} \quad x = 0, 1$$

Suppose that we do not know whether or not the coin is fair; Let θ be the probability of a head expressed as a proportion (between 0 and 1).

$$p(x) = \theta^x (1-\theta)^{1-x} \quad \text{for} \quad x = 0, 1$$

PDF

A probability density function (pdf), is a function associated with a continuous random variable

Areas under pdfs correspond to probabilities for that random variable

To be a valid pdf, a function must satisfy

1. It must be larger than or equal to zero everywhere.
2. The total area under it must be one.

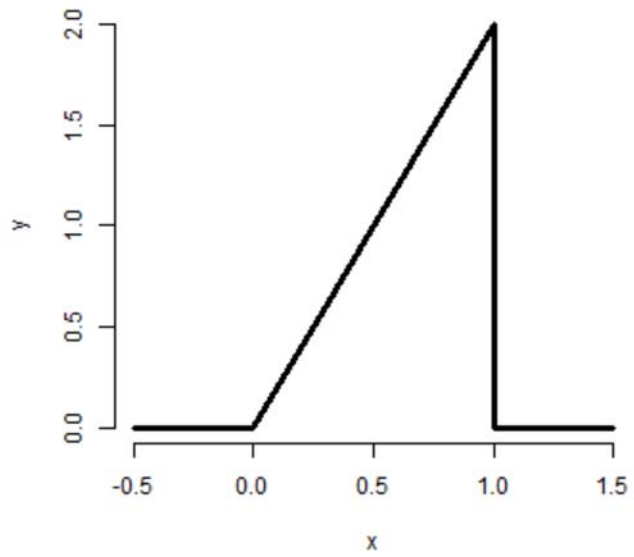
Example

Suppose that the proportion of help calls that get addressed in a random day by a help line is given by

$$f(x) = \begin{cases} 2x & \text{for} \quad 0 < x < 1 \\ 0 & \text{otherwise} \end{cases}$$

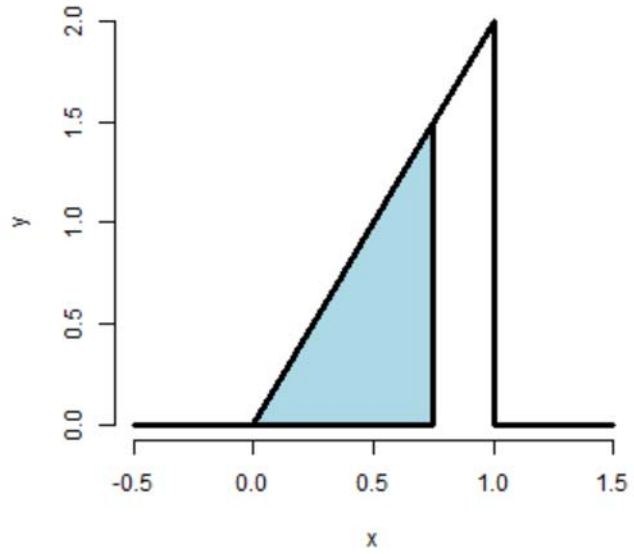
Is this a mathematically valid density?

```r
x <- c(-0.5, 0, 1, 1, 1.5)
y <- c(0, 0, 2, 0, 0)
plot(x, y, lwd = 3, frame = FALSE, type = "l")
```



Example continued

What is the probability that 75% or fewer of calls get addressed?



```
1.5 * 0.75/2
```

```
## [1] 0.5625
```

```
pbeta(0.75, 2, 1)
```

```
## [1] 0.5625
```

CDF and survival function

Certain areas are so useful, we give them names

- The **cumulative distribution function** (CDF) of a random variable, X, returns the probability that the random variable is less than or equal to the value x

$$F(x) = P(X \leq x)$$

  (This definition applies regardless of whether X is discrete or continuous.)

- The **survival function** of a random variable X is defined as the probability that the random variable is greater than the value x

$$S(x) = P(X > x)$$

- Notice that $S(x) = 1 - F(x)$

Example

What are the survival function and CDF from the density considered before?

For $1 \geq x \geq 0$

$$F(x) = P(X \leq x) = \frac{1}{2} \, Base \times Height = \frac{1}{2} \, (x) \times (2x) = x^2$$

$$S(x) = 1 - x^2$$

```
pbeta(c(0.4, 0.5, 0.6), 2, 1)
```

```
## [1] 0.16 0.25 0.36
```

Quantiles

You've heard of sample quantiles. If you were the 95th percentile on an exam, you know that 95% of people scored worse than you and 5% scored better. These are sample quantities. Here we define their population analogs.

Definition

The α^{th} **quantile** of a distribution with distribution function F is the point x_α so that

$$F(x_\alpha) = \alpha$$

- A **percentile** is simply a quantile with α expressed as a percent
- The **median** is the 50^{th} percentile

For example

The 95^{th} percentile of a distribution is the point so that:

- the probability that a random variable drawn from the population is less is 95%
- the probability that a random variable drawn from the population is more is 5%

Example

What is the median of the distribution that we were working with before?

- We want to solve $0.5 = F(x) = x^2$
- Resulting in the solution

```
sqrt(0.5)
```

```
## [1] 0.7071
```

- Therefore, about 0.7071 of calls being answered on a random day is the median.

Example continued

R can approximate quantiles for you for common distributions

```
qbeta(0.5, 2, 1)
```

```
## [1] 0.7071
```

Summary

- You might be wondering at this point "I've heard of a median before, it didn't require integration. Where's the data?"
- We're referring to are **population quantities**. Therefore, the median being discussed is the **population median**.
- A probability model connects the data to the population using assumptions.
- Therefore the median we're discussing is the **estimand**, the sample median will be the **estimator**

Probability

Statistical Inference

Brian Caffo, Jeff Leek, Roger Peng
Johns Hopkins Bloomberg School of Public Health

Probability

- In these slides we will cover the basics of probability at low enough level to have a basic understanding for the rest of the series
- For a more complete treatment see the class Mathematical Biostatistics Boot Camp 1
  - Youtube: www.youtube.com/playlist?list=PLpl-gQkQivXhk6qSyiNj51qamjAtZISJ-
  - Coursera: www.coursera.org/course/biostats
  - Git: http://github.com/bcaffo/Caffo-Coursera

Probability

Given a random experiment (say rolling a die) a probability measure is a population quantity that summarizes the randomness.

Specifically, probability takes a possible outcome from the expertiment and:

- assigns it a number between 0 and 1
- so that the probability that something occurs is 1 (the die must be rolled) and
- so that the probability of the union of any two sets of outcomes that have nothing in common (mutually exclusive) is the sum of their respective probabilities.

The Russian mathematician Kolmogorov formalized these rules.

Rules probability must follow

- The probability that nothing occurs is 0
- The probability that something occurs is 1
- The probability of something is 1 minus the probability that the opposite occurs
- The probability of at least one of two (or more) things that can not simultaneously occur (mutually exclusive) is the sum of their respective probabilities
- If an event A implies the occurrence of event B, then the probability of A occurring is less than the probability that B occurs
- For any two events the probability that at least one occurs is the sum of their probabilities minus their intersection.

Example

The National Sleep Foundation (www.sleepfoundation.org) reports that around 3% of the American population has sleep apnea. They also report that around 10% of the North American and European population has restless leg syndrome. Does this imply that 13% of people will have at least one sleep problems of these sorts?

Example continued

Answer: No, the events can simultaneously occur and so are not mutually exclusive. To elaborate let:

If you want to see the mathematics

$$A_1 = \{\text{Person has sleep apnea}\}$$
$$A_2 = \{\text{Person has RLS}\}$$

Then

$$\begin{aligned} P(A_1 \cup A_2) &= P(A_1) + P(A_2) - P(A_1 \cap A_2) \\ &= 0.13 - \text{Probability of having both} \end{aligned}$$

Likely, some fraction of the population has both.

Going further

Probability calculus is useful for understanding the rules that probabilities must follow.

However, we need ways to model and think about probabilities for numeric outcomes of experiments (broadly defined).

Densities and mass functions for random variables are the best starting point for this.

Remember, everything we're talking about up to at this point is a population quantity not a statement about what occurs in the data.

- We're going with this is: use the data to estimate properties of the population.

Random variables

- A **random variable** is a numerical outcome of an experiment.
- The random variables that we study will come in two varieties, **discrete** or **continuous**.
- Discrete random variable are random variables that take on only a countable number of possibilities and we talk about the probability that they take specific values
- Continuous random variable can conceptually take any value on the real line or some subset of the real line and we talk about the probability that they line within some range

Examples of variables that can be thought of as random variables

Experiments that we use for intuition and building context

- The $(0-1)$ outcome of the flip of a coin
- The outcome from the roll of a die

Specific instances of treating variables as if random

- The web site traffic on a given day
- The BMI of a subject four years after a baseline measurement
- The hypertension status of a subject randomly drawn from a population
- The number of people who click on an ad
- Intelligence quotients for a sample of children

PMF

A probability mass function evaluated at a value corresponds to the probability that a random variable takes that value. To be a valid pmf a function, p, must satisfy

1. It must always be larger than or equal to 0.
2. The sum of the possible values that the random variable can take has to add up to one.

Example

Let X be the result of a coin flip where $X = 0$ represents tails and $X = 1$ represents heads.

$$p(x) = (1/2)^x (1/2)^{1-x} \quad \text{for} \quad x = 0, 1$$

Suppose that we do not know whether or not the coin is fair; Let θ be the probability of a head expressed as a proportion (between 0 and 1).

$$p(x) = \theta^x (1-\theta)^{1-x} \quad \text{for} \quad x = 0, 1$$

PDF

A probability density function (pdf), is a function associated with a continuous random variable

Areas under pdfs correspond to probabilities for that random variable

To be a valid pdf, a function must satisfy

1. It must be larger than or equal to zero everywhere.
2. The total area under it must be one.

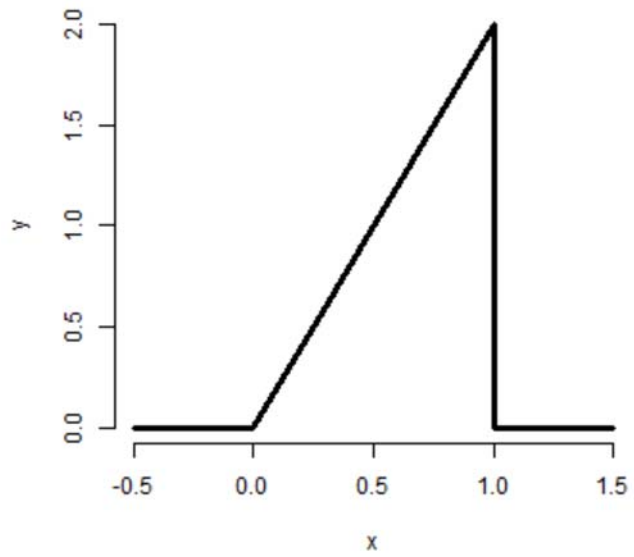
Example

Suppose that the proportion of help calls that get addressed in a random day by a help line is given by

$$f(x) = \begin{cases} 2x & \text{for} \quad 0 < x < 1 \\ 0 & \text{otherwise} \end{cases}$$

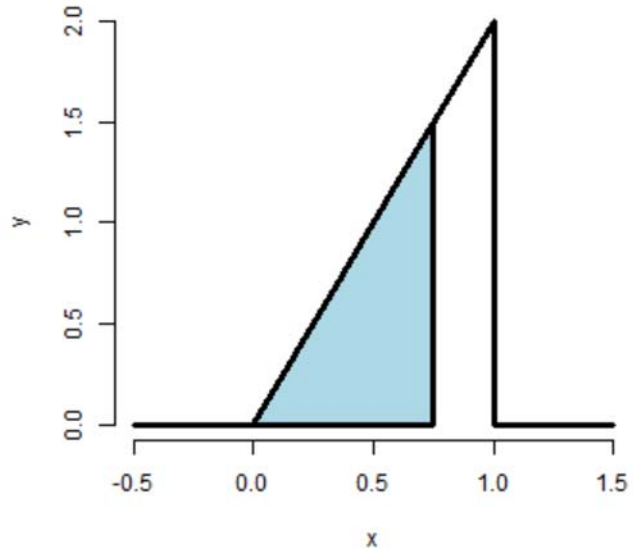
Is this a mathematically valid density?

```r
x <- c(-0.5, 0, 1, 1, 1.5)
y <- c(0, 0, 2, 0, 0)
plot(x, y, lwd = 3, frame = FALSE, type = "l")
```



Example continued

What is the probability that 75% or fewer of calls get addressed?



```
1.5 * 0.75/2
```

```
## [1] 0.5625
```

```
pbeta(0.75, 2, 1)
```

```
## [1] 0.5625
```

CDF and survival function

Certain areas are so useful, we give them names

- The **cumulative distribution function** (CDF) of a random variable, X, returns the probability that the random variable is less than or equal to the value x

$$F(x) = P(X \leq x)$$

  (This definition applies regardless of whether X is discrete or continuous.)

- The **survival function** of a random variable X is defined as the probability that the random variable is greater than the value x

$$S(x) = P(X > x)$$

- Notice that $S(x) = 1 - F(x)$

Example

What are the survival function and CDF from the density considered before?

For $1 \geq x \geq 0$

$$F(x) = P(X \leq x) = \frac{1}{2} \, Base \times Height = \frac{1}{2} \, (x) \times (2x) = x^2$$

$$S(x) = 1 - x^2$$

```
pbeta(c(0.4, 0.5, 0.6), 2, 1)
```

```
## [1] 0.16 0.25 0.36
```

Quantiles

You've heard of sample quantiles. If you were the 95th percentile on an exam, you know that 95% of people scored worse than you and 5% scored better. These are sample quantities. Here we define their population analogs.

Definition

The α^{th} **quantile** of a distribution with distribution function F is the point x_α so that

$$F(x_\alpha) = \alpha$$

- A **percentile** is simply a quantile with α expressed as a percent
- The **median** is the 50^{th} percentile

For example

The 95^{th} percentile of a distribution is the point so that:

- the probability that a random variable drawn from the population is less is 95%
- the probability that a random variable drawn from the population is more is 5%

Example

What is the median of the distribution that we were working with before?

- We want to solve $0.5 = F(x) = x^2$
- Resulting in the solution

```
sqrt(0.5)
```

```
## [1] 0.7071
```

- Therefore, about 0.7071 of calls being answered on a random day is the median.

Example continued

R can approximate quantiles for you for common distributions

```
qbeta(0.5, 2, 1)
```

```
## [1] 0.7071
```

Summary

- You might be wondering at this point "I've heard of a median before, it didn't require integration. Where's the data?"
- We're referring to are **population quantities**. Therefore, the median being discussed is the **population median**.
- A probability model connects the data to the population using assumptions.
- Therefore the median we're discussing is the **estimand**, the sample median will be the **estimator**

Conditional Probability

Statistical Inference

Brian Caffo, Jeff Leek, Roger Peng
Johns Hopkins Bloomberg School of Public Health

Conditional probability, motivation

- The probability of getting a one when rolling a (standard) die is usually assumed to be one sixth
- Suppose you were given the extra information that the die roll was an odd number (hence 1, 3 or 5)
- *conditional on this new information*, the probability of a one is now one third

Conditional probability, definition

- Let B be an event so that $P(B) > 0$
- Then the conditional probability of an event A given that B has occurred is

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

- Notice that if A and B are independent (defined later in the lecture), then

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

Example

- Consider our die roll example
- $B = \{1, 3, 5\}$
- $A = \{1\}$

$$
\begin{aligned}
P(\text{one given that roll is odd}) &= P(A \mid B) \\
&= \frac{P(A \cap B)}{P(B)} \\
&= \frac{P(A)}{P(B)} \\
&= \frac{1/6}{3/6} = \frac{1}{3}
\end{aligned}
$$

Bayes' rule

Baye's rule allows us to reverse the conditioning set provided that we know some marginal probabilities

$$P(B \mid A) = \frac{P(A \mid B)P(B)}{P(A \mid B)P(B) + P(A \mid B^c)P(B^c)} \, .$$

Diagnostic tests

- Let $+$ and $-$ be the events that the result of a diagnostic test is positive or negative respectively
- Let D and D^c be the event that the subject of the test has or does not have the disease respectively
- The **sensitivity** is the probability that the test is positive given that the subject actually has the disease, $P(+ \mid D)$
- The **specificity** is the probability that the test is negative given that the subject does not have the disease, $P(- \mid D^c)$

More definitions

- The **positive predictive value** is the probability that the subject has the disease given that the test is positive, $P(D \mid +)$
- The **negative predictive value** is the probability that the subject does not have the disease given that the test is negative, $P(D^c \mid -)$
- The **prevalence of the disease** is the marginal probability of disease, $P(D)$

More definitions

- The **diagnostic likelihood ratio of a positive test**, labeled DLR_+, is $P(+ \mid D)/P(+ \mid D^c)$, which is the

$$sensitivity/(1 - specificity)$$

- The **diagnostic likelihood ratio of a negative test**, labeled DLR_-, is $P(- \mid D)/P(- \mid D^c)$, which is the

$$(1 - sensitivity)/specificity$$

Example

- A study comparing the efficacy of HIV tests, reports on an experiment which concluded that HIV antibody tests have a sensitivity of 99.7% and a specificity of 98.5%
- Suppose that a subject, from a population with a .1% prevalence of HIV, receives a positive test result. What is the positive predictive value?
- Mathematically, we want $P(D \mid +)$ given the sensitivity, $P(+ \mid D) = .997$, the specificity, $P(- \mid D^c) = .985$, and the prevalence $P(D) = .001$

Using Bayes' formula

$$
\begin{aligned}
P(D \mid +) &= \frac{P(+ \mid D)P(D)}{P(+ \mid D)P(D) + P(+ \mid D^c)P(D^c)} \\
&= \frac{P(+ \mid D)P(D)}{P(+ \mid D)P(D) + \{1 - P(- \mid D^c)\}\{1 - P(D)\}} \\
&= \frac{.997 \times .001}{.997 \times .001 + .015 \times .999} \\
&= .062
\end{aligned}
$$

- In this population a positive test result only suggests a 6% probability that the subject has the disease
- (The positive predictive value is 6% for this test)

More on this example

- The low positive predictive value is due to low prevalence of disease and the somewhat modest specificity
- Suppose it was known that the subject was an intravenous drug user and routinely had intercourse with an HIV infected partner
- Notice that the evidence implied by a positive test result does not change because of the prevalence of disease in the subject's population, only our interpretation of that evidence changes

Likelihood ratios

- Using Bayes rule, we have

$$P(D \mid +) = \frac{P(+ \mid D)P(D)}{P(+ \mid D)P(D) + P(+ \mid D^c)P(D^c)}$$

and

$$P(D^c \mid +) = \frac{P(+ \mid D^c)P(D^c)}{P(+ \mid D)P(D) + P(+ \mid D^c)P(D^c)} .$$

Likelihood ratios

- Therefore

$$\frac{P(D \mid +)}{P(D^c \mid +)} = \frac{P(+ \mid D)}{P(+ \mid D^c)} \times \frac{P(D)}{P(D^c)}$$

  ie

$$\text{post-test odds of } D = DLR_+ \times \text{pre-test odds of } D$$

- Similarly, DLR_- relates the decrease in the odds of the disease after a negative test result to the odds of disease prior to the test.

HIV example revisited

- Suppose a subject has a positive HIV test
- $DLR_+ = .997/(1 - .985) \approx 66$
- The result of the positive test is that the odds of disease is now 66 times the pretest odds
- Or, equivalently, the hypothesis of disease is 66 times more supported by the data than the hypothesis of no disease

HIV example revisited

- Suppose that a subject has a negative test result
- $DLR_{-} = (1 - .997)/.985 \approx .003$
- Therefore, the post-test odds of disease is now $.3\%$ of the pretest odds given the negative test.
- Or, the hypothesis of disease is supported $.003$ times that of the hypothesis of absence of disease given the negative test result

Independence

- Two events A and B are **independent** if

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

- Equivalently if $P(A \mid B) = P(A)$
- Two random variables, X and Y are independent if for any two sets A and B

$$P([X \in A] \cap [Y \in B]) = P(X \in A)P(Y \in B)$$

- If A is independent of B then
  - A^c is independent of B
  - A is independent of B^c
  - A^c is independent of B^c

Example

- What is the probability of getting two consecutive heads?
- $A = \{\text{Head on flip } 1\}$ ~ $P(A) = .5$
- $B = \{\text{Head on flip } 2\}$ ~ $P(B) = .5$
- $A \cap B = \{\text{Head on flips } 1 \text{ and } 2\}$
- $P(A \cap B) = P(A)P(B) = .5 \times .5 = .25$

Example

- Volume 309 of Science reports on a physician who was on trial for expert testimony in a criminal trial
- Based on an estimated prevalence of sudden infant death syndrome of 1 out of $8,543$, the physician testified that that the probability of a mother having two children with SIDS was $\left(\frac{1}{8,543}\right)^2$
- The mother on trial was convicted of murder

Example: continued

- Relevant to this discussion, the principal mistake was to *assume* that the events of having SIDs within a family are independent
- That is, $P(A_1 \cap A_2)$ is not necessarily equal to $P(A_1)P(A_2)$
- Biological processes that have a believed genetic or familiar environmental component, of course, tend to be dependent within families
- (There are many other statistical points of discussion for this case.)

IID random variables

- Random variables are said to be iid if they are independent and identically distributed
  - Independent: statistically unrelated from one and another
  - Identically distributed: all having been drawn from the same population distribution
- iid random variables are the default model for random samples
- Many of the important theories of statistics are founded on assuming that variables are iid
- Assuming a random sample and iid will be the default starting point of inference for this class

Conditional Probability

Statistical Inference

Brian Caffo, Jeff Leek, Roger Peng
Johns Hopkins Bloomberg School of Public Health

Conditional probability, motivation

- The probability of getting a one when rolling a (standard) die is usually assumed to be one sixth
- Suppose you were given the extra information that the die roll was an odd number (hence 1, 3 or 5)
- *conditional on this new information*, the probability of a one is now one third

Conditional probability, definition

- Let B be an event so that $P(B) > 0$
- Then the conditional probability of an event A given that B has occurred is

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

- Notice that if A and B are independent (defined later in the lecture), then

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

Example

- Consider our die roll example
- $B = \{1, 3, 5\}$
- $A = \{1\}$

$$
\begin{aligned}
P(\text{one given that roll is odd}) &= P(A \mid B) \\
&= \frac{P(A \cap B)}{P(B)} \\
&= \frac{P(A)}{P(B)} \\
&= \frac{1/6}{3/6} = \frac{1}{3}
\end{aligned}
$$

Bayes' rule

Baye's rule allows us to reverse the conditioning set provided that we know some marginal probabilities

$$P(B \mid A) = \frac{P(A \mid B)P(B)}{P(A \mid B)P(B) + P(A \mid B^c)P(B^c)} \, .$$

Diagnostic tests

- Let $+$ and $-$ be the events that the result of a diagnostic test is positive or negative respectively
- Let D and D^c be the event that the subject of the test has or does not have the disease respectively
- The **sensitivity** is the probability that the test is positive given that the subject actually has the disease, $P(+ \mid D)$
- The **specificity** is the probability that the test is negative given that the subject does not have the disease, $P(- \mid D^c)$

More definitions

- The **positive predictive value** is the probability that the subject has the disease given that the test is positive, $P(D \mid +)$
- The **negative predictive value** is the probability that the subject does not have the disease given that the test is negative, $P(D^c \mid -)$
- The **prevalence of the disease** is the marginal probability of disease, $P(D)$

More definitions

- The **diagnostic likelihood ratio of a positive test**, labeled DLR_+, is $P(+ \mid D)/P(+ \mid D^c)$, which is the

$$sensitivity/(1 - specificity)$$

- The **diagnostic likelihood ratio of a negative test**, labeled DLR_-, is $P(- \mid D)/P(- \mid D^c)$, which is the

$$(1 - sensitivity)/specificity$$

Example

- A study comparing the efficacy of HIV tests, reports on an experiment which concluded that HIV antibody tests have a sensitivity of 99.7% and a specificity of 98.5%
- Suppose that a subject, from a population with a .1% prevalence of HIV, receives a positive test result. What is the positive predictive value?
- Mathematically, we want $P(D \mid +)$ given the sensitivity, $P(+ \mid D) = .997$, the specificity, $P(- \mid D^c) = .985$, and the prevalence $P(D) = .001$

Using Bayes' formula

$$
\begin{aligned}
P(D \mid +) &= \frac{P(+ \mid D)P(D)}{P(+ \mid D)P(D) + P(+ \mid D^c)P(D^c)} \\
&= \frac{P(+ \mid D)P(D)}{P(+ \mid D)P(D) + \{1 - P(- \mid D^c)\}\{1 - P(D)\}} \\
&= \frac{.997 \times .001}{.997 \times .001 + .015 \times .999} \\
&= .062
\end{aligned}
$$

- In this population a positive test result only suggests a 6% probability that the subject has the disease
- (The positive predictive value is 6% for this test)

More on this example

- The low positive predictive value is due to low prevalence of disease and the somewhat modest specificity
- Suppose it was known that the subject was an intravenous drug user and routinely had intercourse with an HIV infected partner
- Notice that the evidence implied by a positive test result does not change because of the prevalence of disease in the subject's population, only our interpretation of that evidence changes

Likelihood ratios

- Using Bayes rule, we have

$$P(D \mid +) = \frac{P(+ \mid D)P(D)}{P(+ \mid D)P(D) + P(+ \mid D^c)P(D^c)}$$

and

$$P(D^c \mid +) = \frac{P(+ \mid D^c)P(D^c)}{P(+ \mid D)P(D) + P(+ \mid D^c)P(D^c)} .$$

Likelihood ratios

- Therefore

$$\frac{P(D \mid +)}{P(D^c \mid +)} = \frac{P(+ \mid D)}{P(+ \mid D^c)} \times \frac{P(D)}{P(D^c)}$$

  ie

$$\text{post-test odds of } D = DLR_+ \times \text{pre-test odds of } D$$

- Similarly, DLR_- relates the decrease in the odds of the disease after a negative test result to the odds of disease prior to the test.

HIV example revisited

- Suppose a subject has a positive HIV test
- $DLR_+ = .997/(1 - .985) \approx 66$
- The result of the positive test is that the odds of disease is now 66 times the pretest odds
- Or, equivalently, the hypothesis of disease is 66 times more supported by the data than the hypothesis of no disease

HIV example revisited

- Suppose that a subject has a negative test result
- $DLR_{-} = (1 - .997)/.985 \approx .003$
- Therefore, the post-test odds of disease is now $.3\%$ of the pretest odds given the negative test.
- Or, the hypothesis of disease is supported $.003$ times that of the hypothesis of absence of disease given the negative test result

Independence

- Two events A and B are **independent** if

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

- Equivalently if $P(A \mid B) = P(A)$
- Two random variables, X and Y are independent if for any two sets A and B

$$P([X \in A] \cap [Y \in B]) = P(X \in A)P(Y \in B)$$

- If A is independent of B then
  - A^c is independent of B
  - A is independent of B^c
  - A^c is independent of B^c

Example

- What is the probability of getting two consecutive heads?
- $A = \{\text{Head on flip } 1\}$ ~ $P(A) = .5$
- $B = \{\text{Head on flip } 2\}$ ~ $P(B) = .5$
- $A \cap B = \{\text{Head on flips } 1 \text{ and } 2\}$
- $P(A \cap B) = P(A)P(B) = .5 \times .5 = .25$

Example

- Volume 309 of Science reports on a physician who was on trial for expert testimony in a criminal trial
- Based on an estimated prevalence of sudden infant death syndrome of 1 out of $8,543$, the physician testified that that the probability of a mother having two children with SIDS was $\left(\frac{1}{8,543}\right)^2$
- The mother on trial was convicted of murder

Example: continued

- Relevant to this discussion, the principal mistake was to *assume* that the events of having SIDs within a family are independent
- That is, $P(A_1 \cap A_2)$ is not necessarily equal to $P(A_1)P(A_2)$
- Biological processes that have a believed genetic or familiar environmental component, of course, tend to be dependent within families
- (There are many other statistical points of discussion for this case.)

IID random variables

- Random variables are said to be iid if they are independent and identically distributed
  - Independent: statistically unrelated from one and another
  - Identically distributed: all having been drawn from the same population distribution
- iid random variables are the default model for random samples
- Many of the important theories of statistics are founded on assuming that variables are iid
- Assuming a random sample and iid will be the default starting point of inference for this class

Conditional Probability

Statistical Inference

Brian Caffo, Jeff Leek, Roger Peng
Johns Hopkins Bloomberg School of Public Health

Conditional probability, motivation

- The probability of getting a one when rolling a (standard) die is usually assumed to be one sixth
- Suppose you were given the extra information that the die roll was an odd number (hence 1, 3 or 5)
- *conditional on this new information*, the probability of a one is now one third

Conditional probability, definition

- Let B be an event so that $P(B) > 0$
- Then the conditional probability of an event A given that B has occurred is

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

- Notice that if A and B are independent (defined later in the lecture), then

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

Example

- Consider our die roll example
- $B = \{1, 3, 5\}$
- $A = \{1\}$

$$
\begin{aligned}
P(\text{one given that roll is odd}) &= P(A \mid B) \\
&= \frac{P(A \cap B)}{P(B)} \\
&= \frac{P(A)}{P(B)} \\
&= \frac{1/6}{3/6} = \frac{1}{3}
\end{aligned}
$$

Bayes' rule

Baye's rule allows us to reverse the conditioning set provided that we know some marginal probabilities

$$P(B \mid A) = \frac{P(A \mid B)P(B)}{P(A \mid B)P(B) + P(A \mid B^c)P(B^c)} \, .$$

Diagnostic tests

- Let $+$ and $-$ be the events that the result of a diagnostic test is positive or negative respectively
- Let D and D^c be the event that the subject of the test has or does not have the disease respectively
- The **sensitivity** is the probability that the test is positive given that the subject actually has the disease, $P(+ \mid D)$
- The **specificity** is the probability that the test is negative given that the subject does not have the disease, $P(- \mid D^c)$

More definitions

- The **positive predictive value** is the probability that the subject has the disease given that the test is positive, $P(D \mid +)$
- The **negative predictive value** is the probability that the subject does not have the disease given that the test is negative, $P(D^c \mid -)$
- The **prevalence of the disease** is the marginal probability of disease, $P(D)$

More definitions

- The **diagnostic likelihood ratio of a positive test**, labeled DLR_+, is $P(+ \mid D)/P(+ \mid D^c)$, which is the

$$sensitivity/(1 - specificity)$$

- The **diagnostic likelihood ratio of a negative test**, labeled DLR_-, is $P(- \mid D)/P(- \mid D^c)$, which is the

$$(1 - sensitivity)/specificity$$

Example

- A study comparing the efficacy of HIV tests, reports on an experiment which concluded that HIV antibody tests have a sensitivity of 99.7% and a specificity of 98.5%
- Suppose that a subject, from a population with a .1% prevalence of HIV, receives a positive test result. What is the positive predictive value?
- Mathematically, we want $P(D \mid +)$ given the sensitivity, $P(+ \mid D) = .997$, the specificity, $P(- \mid D^c) = .985$, and the prevalence $P(D) = .001$

Using Bayes' formula

$$
\begin{aligned}
P(D \mid +) &= \frac{P(+ \mid D)P(D)}{P(+ \mid D)P(D) + P(+ \mid D^c)P(D^c)} \\
&= \frac{P(+ \mid D)P(D)}{P(+ \mid D)P(D) + \{1 - P(- \mid D^c)\}\{1 - P(D)\}} \\
&= \frac{.997 \times .001}{.997 \times .001 + .015 \times .999} \\
&= .062
\end{aligned}
$$

- In this population a positive test result only suggests a 6% probability that the subject has the disease
- (The positive predictive value is 6% for this test)

More on this example

- The low positive predictive value is due to low prevalence of disease and the somewhat modest specificity
- Suppose it was known that the subject was an intravenous drug user and routinely had intercourse with an HIV infected partner
- Notice that the evidence implied by a positive test result does not change because of the prevalence of disease in the subject's population, only our interpretation of that evidence changes

Likelihood ratios

- Using Bayes rule, we have

$$P(D \mid +) = \frac{P(+ \mid D)P(D)}{P(+ \mid D)P(D) + P(+ \mid D^c)P(D^c)}$$

and

$$P(D^c \mid +) = \frac{P(+ \mid D^c)P(D^c)}{P(+ \mid D)P(D) + P(+ \mid D^c)P(D^c)} .$$

Likelihood ratios

- Therefore

$$\frac{P(D \mid +)}{P(D^c \mid +)} = \frac{P(+ \mid D)}{P(+ \mid D^c)} \times \frac{P(D)}{P(D^c)}$$

  ie

$$\text{post-test odds of } D = DLR_+ \times \text{pre-test odds of } D$$

- Similarly, DLR_- relates the decrease in the odds of the disease after a negative test result to the odds of disease prior to the test.

HIV example revisited

- Suppose a subject has a positive HIV test
- $DLR_+ = .997/(1 - .985) \approx 66$
- The result of the positive test is that the odds of disease is now 66 times the pretest odds
- Or, equivalently, the hypothesis of disease is 66 times more supported by the data than the hypothesis of no disease

HIV example revisited

- Suppose that a subject has a negative test result
- $DLR_{-} = (1 - .997)/.985 \approx .003$
- Therefore, the post-test odds of disease is now $.3\%$ of the pretest odds given the negative test.
- Or, the hypothesis of disease is supported $.003$ times that of the hypothesis of absence of disease given the negative test result

Independence

- Two events A and B are **independent** if

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

- Equivalently if $P(A \mid B) = P(A)$
- Two random variables, X and Y are independent if for any two sets A and B

$$P([X \in A] \cap [Y \in B]) = P(X \in A)P(Y \in B)$$

- If A is independent of B then
  - A^c is independent of B
  - A is independent of B^c
  - A^c is independent of B^c

Example

- What is the probability of getting two consecutive heads?
- $A = \{\text{Head on flip } 1\}$ ~ $P(A) = .5$
- $B = \{\text{Head on flip } 2\}$ ~ $P(B) = .5$
- $A \cap B = \{\text{Head on flips } 1 \text{ and } 2\}$
- $P(A \cap B) = P(A)P(B) = .5 \times .5 = .25$

Example

- Volume 309 of Science reports on a physician who was on trial for expert testimony in a criminal trial
- Based on an estimated prevalence of sudden infant death syndrome of 1 out of $8,543$, the physician testified that that the probability of a mother having two children with SIDS was $\left(\frac{1}{8,543}\right)^2$
- The mother on trial was convicted of murder

Example: continued

- Relevant to this discussion, the principal mistake was to *assume* that the events of having SIDs within a family are independent
- That is, $P(A_1 \cap A_2)$ is not necessarily equal to $P(A_1)P(A_2)$
- Biological processes that have a believed genetic or familiar environmental component, of course, tend to be dependent within families
- (There are many other statistical points of discussion for this case.)

IID random variables

- Random variables are said to be iid if they are independent and identically distributed
  - Independent: statistically unrelated from one and another
  - Identically distributed: all having been drawn from the same population distribution
- iid random variables are the default model for random samples
- Many of the important theories of statistics are founded on assuming that variables are iid
- Assuming a random sample and iid will be the default starting point of inference for this class

Expected values

Statistical Inference

Brian Caffo, Jeff Leek, Roger Peng
Johns Hopkins Bloomberg School of Public Health

Expected values

- Expected values are useful cor characterizing a distribution
- The mean is a characterization of its center
- The variance and standard deviation are characterizations of how spread out it is
- Our sample expected values (the sample mean and variance) will estimate the population versions

The population mean

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

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

  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, $\{x, p(x)\}$

The sample mean

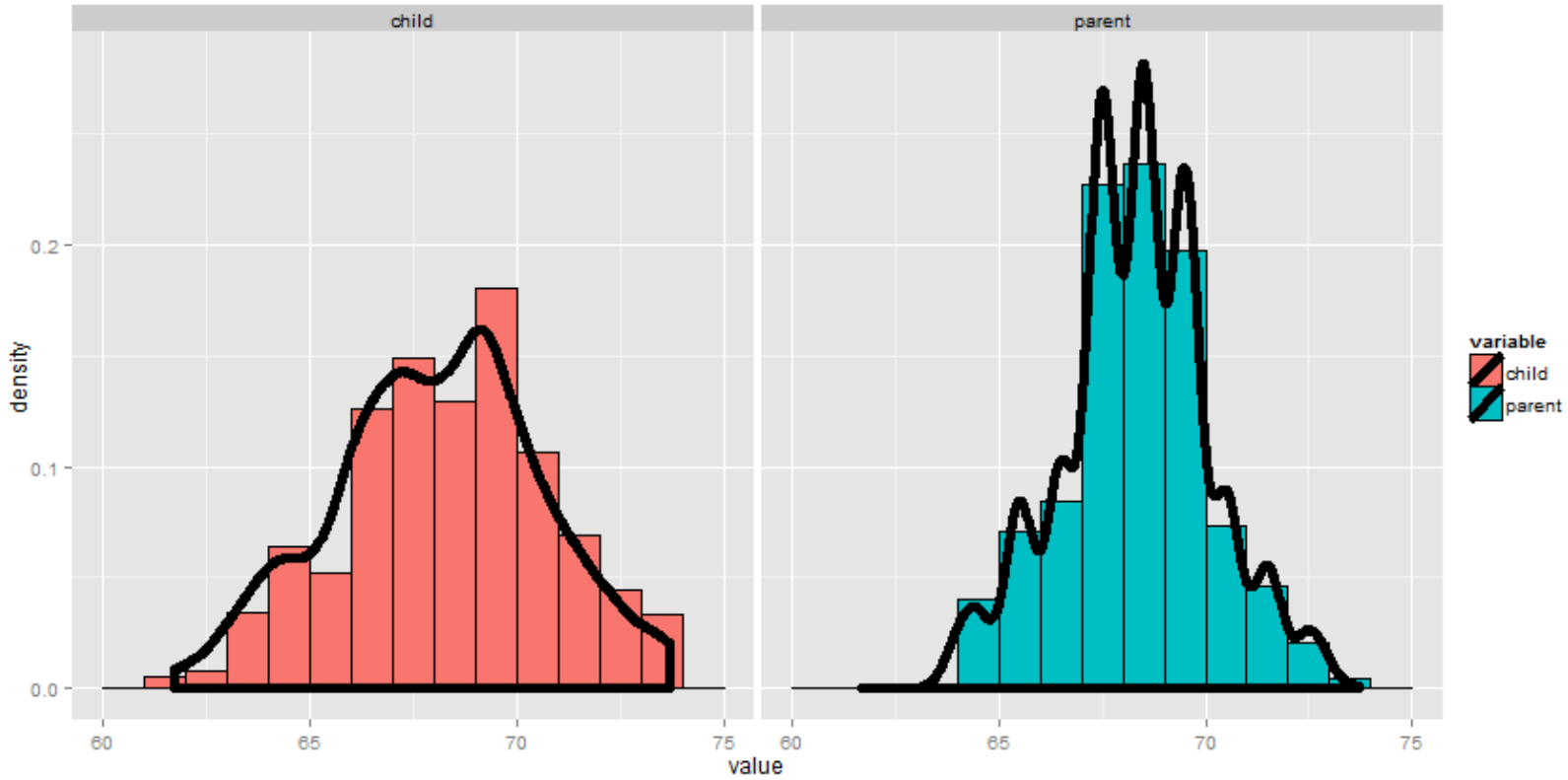
- The sample mean estimates this population mean
- The center of mass of the data is the empirical mean

$$\bar{X} = \sum_{i=1}^{n} x_i p(x_i)$$

where $p(x_i) = 1/n$

Example

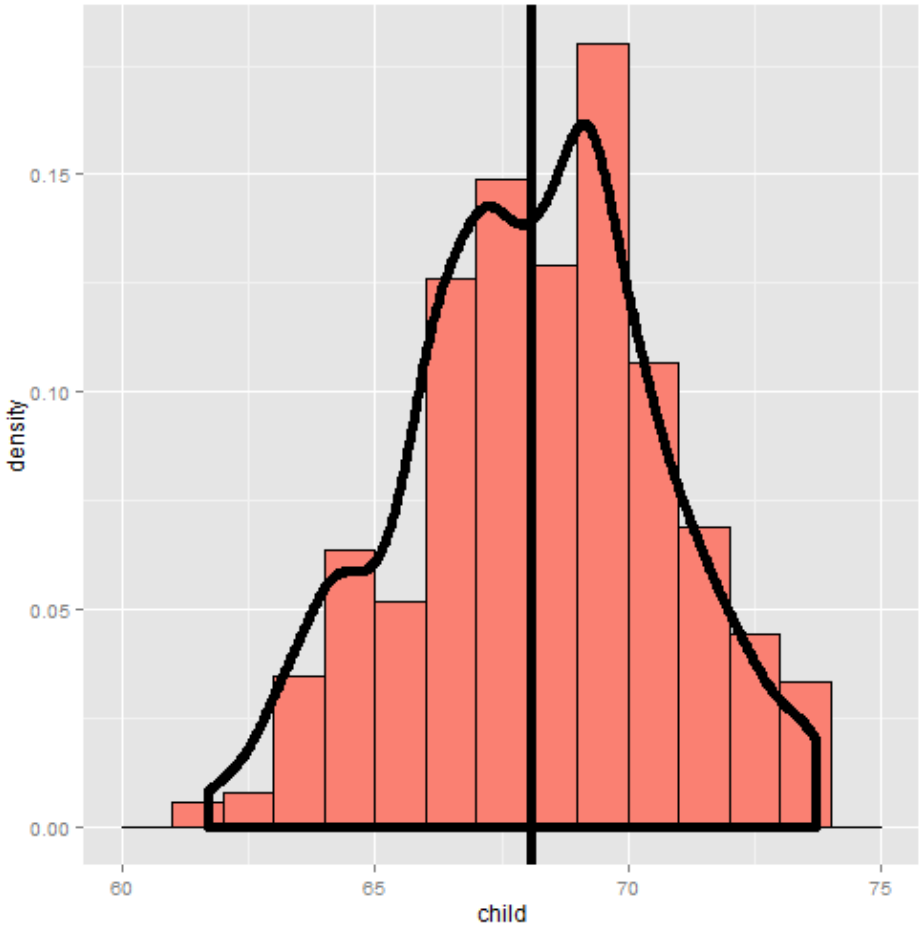
Find the center of mass of the bars



Using manipulate

```
library(manipulate)
myHist <- function(mu){
    g <- ggplot(galton, aes(x = child))
    g <- g + geom_histogram(fill = "salmon",
      binwidth=1, aes(y = ..density..), colour = "black")
    g <- g + geom_density(size = 2)
    g <- g + geom_vline(xintercept = mu, size = 2)
    mse <- round(mean((galton$child - mu)^2), 3)
    g <- g + labs(title = paste('mu = ', mu, ' MSE = ', mse))
    g
}
manipulate(myHist(mu), mu = slider(62, 74, step = 0.5))
```

The center of mass is the empirical mean

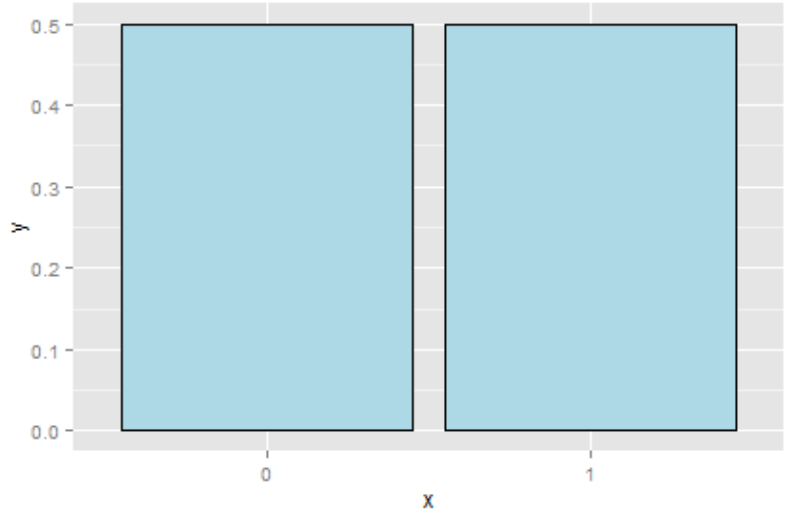


Example of a population mean

- 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 $.5$



What about a biased coin?

- Suppose that a random variable, X, is so that $P(X = 1) = p$ and $P(X = 0) = (1 - p)$
- (This is a biased coin when $p \neq 0.5$)
- What is its expected value?

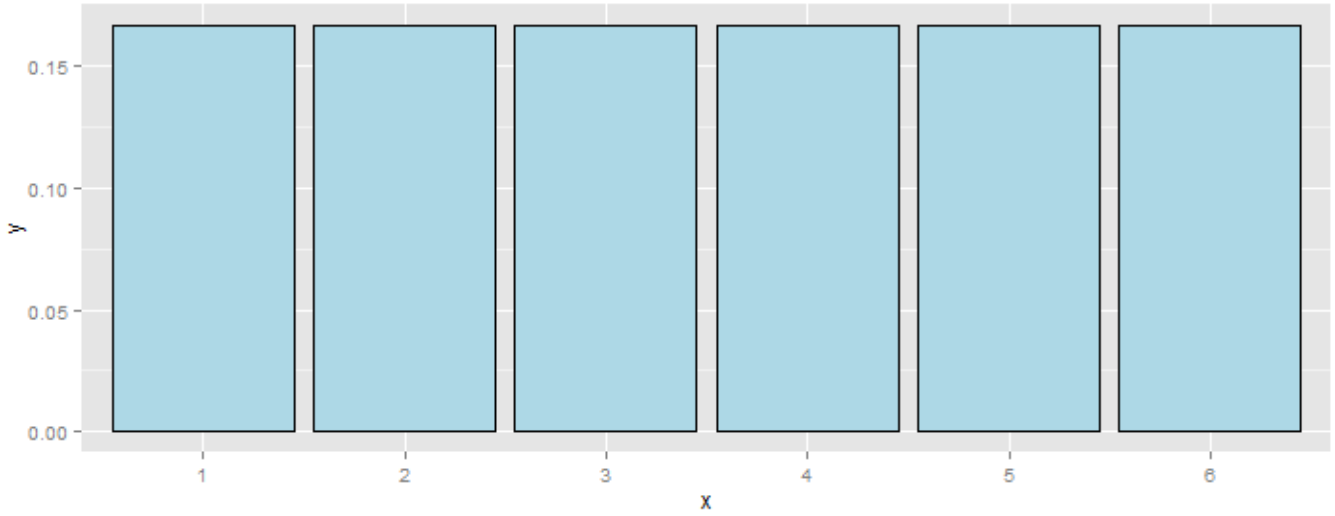
$$E[X] = 0 * (1 - p) + 1 * p = p$$

Example

- Suppose that a die is rolled and X is the number face up
- What is the expected value of X?

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

- Again, the geometric argument makes this answer obvious without calculation.

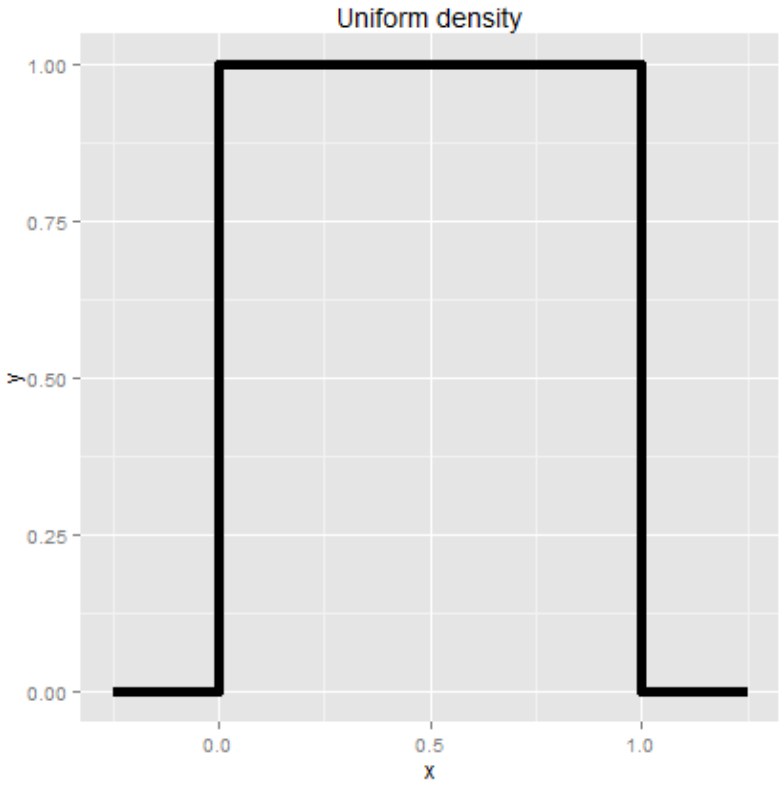


Continuous random variables

- For a continuous random variable, X, with density, f, the expected value is again exactly the center of mass of the density

Example

- Consider a density where $f(x) = 1$ for x between zero and one
- (Is this a valid density?)
- Suppose that X follows this density; what is its expected value?

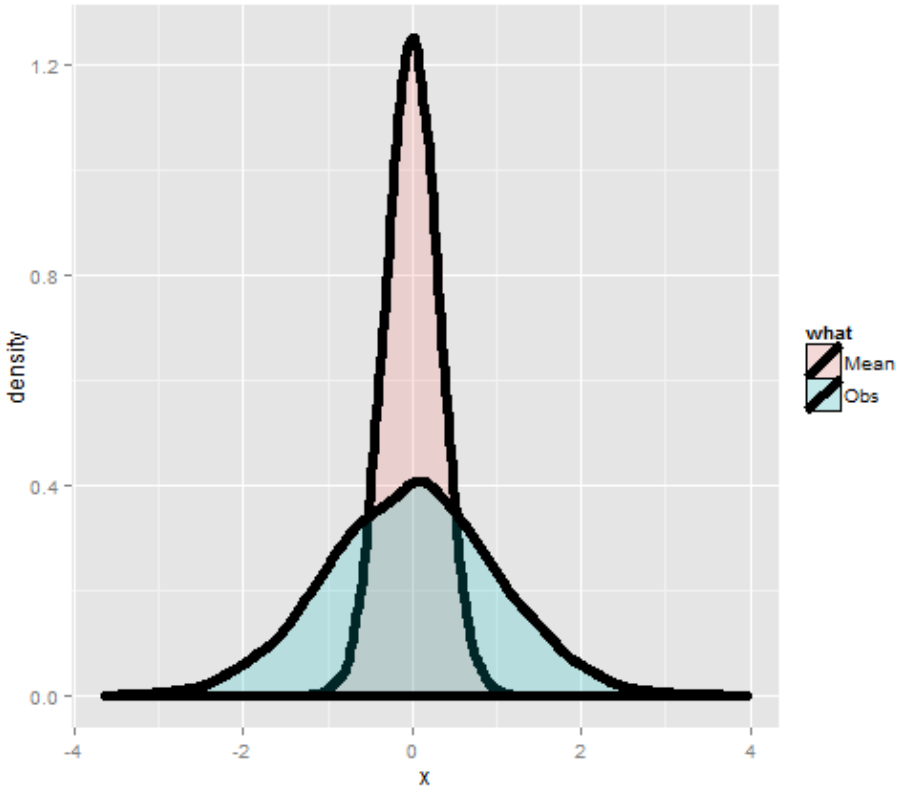


Facts about expected values

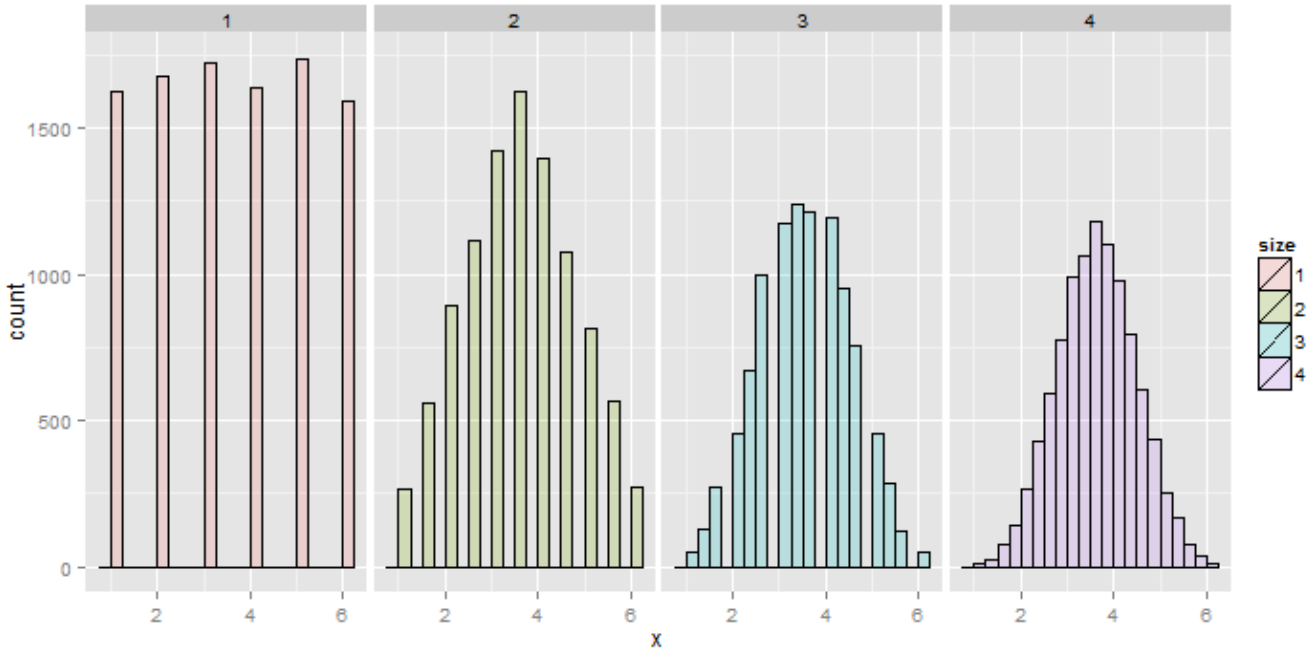
- Recall that expected values are properties of distributions
- Note the average of random variables is itself a random variable and its associated distribution has an expected value
- The center of this distribution is the same as that of the original distribution
- Therefore, the expected value of the **sample mean** is the population mean that it's trying to estimate
- When the expected value of an estimator is what its trying to estimate, we say that the estimator is **unbiased**
- Let's try a simulation experiment

Simulation experiment

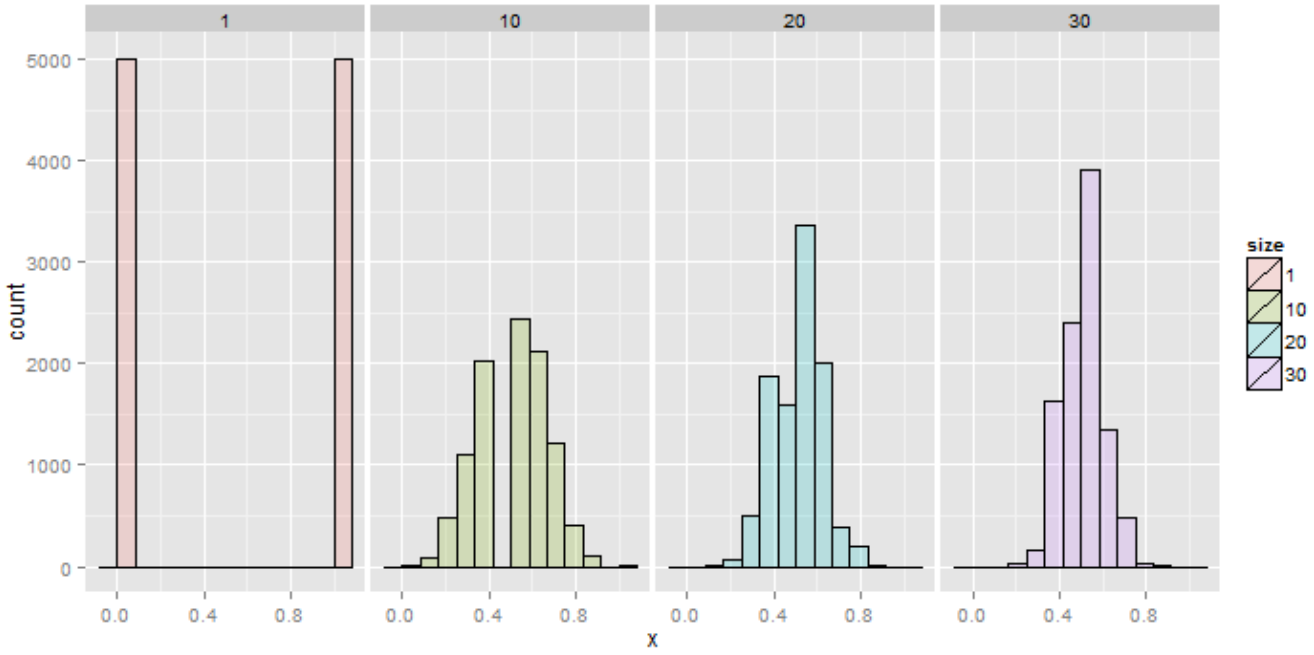
Simulating normals with mean 0 and variance 1 versus averages of 10 normals from the same population



Averages of x die rolls



Averages of x coin flips



Sumarizing what we know

- Expected values are properties of distributions
- The population mean is the center of mass of population
- The sample mean is the center of mass of the observed data
- The sample mean is an estimate of the population mean
- The sample mean is unbiased
  - The population mean of its distribution is the mean that it's trying to estimate
- The more data that goes into the sample mean, the more concentrated its density / mass function is around the population mean

Expected values

Statistical Inference

Brian Caffo, Jeff Leek, Roger Peng
Johns Hopkins Bloomberg School of Public Health

Expected values

- Expected values are useful cor characterizing a distribution
- The mean is a characterization of its center
- The variance and standard deviation are characterizations of how spread out it is
- Our sample expected values (the sample mean and variance) will estimate the population versions

The population mean

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

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

  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, $\{x, p(x)\}$

The sample mean

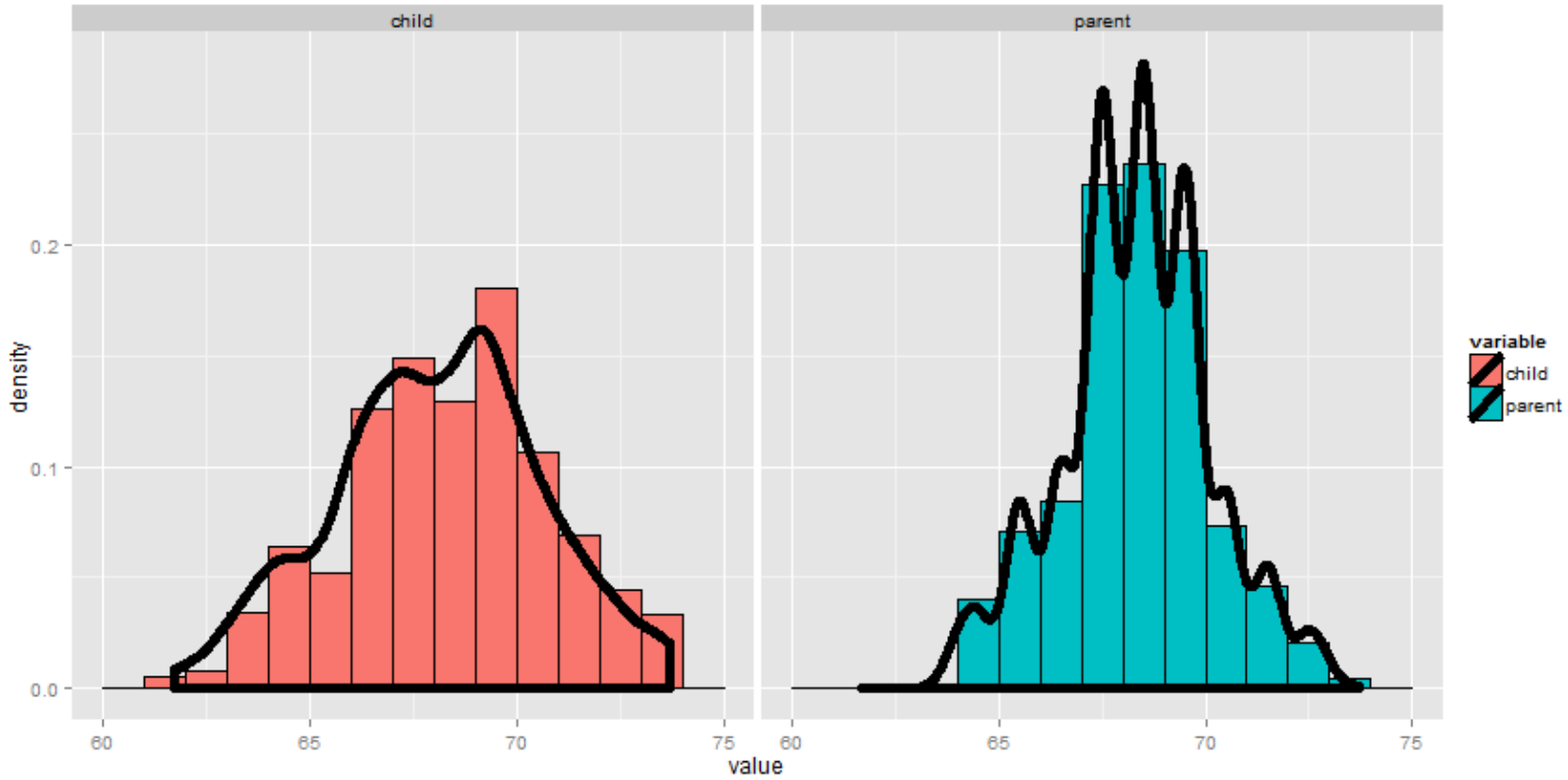
- The sample mean estimates this population mean
- The center of mass of the data is the empirical mean

$$\bar{X} = \sum_{i=1}^{n} x_i p(x_i)$$

where $p(x_i) = 1/n$

Example

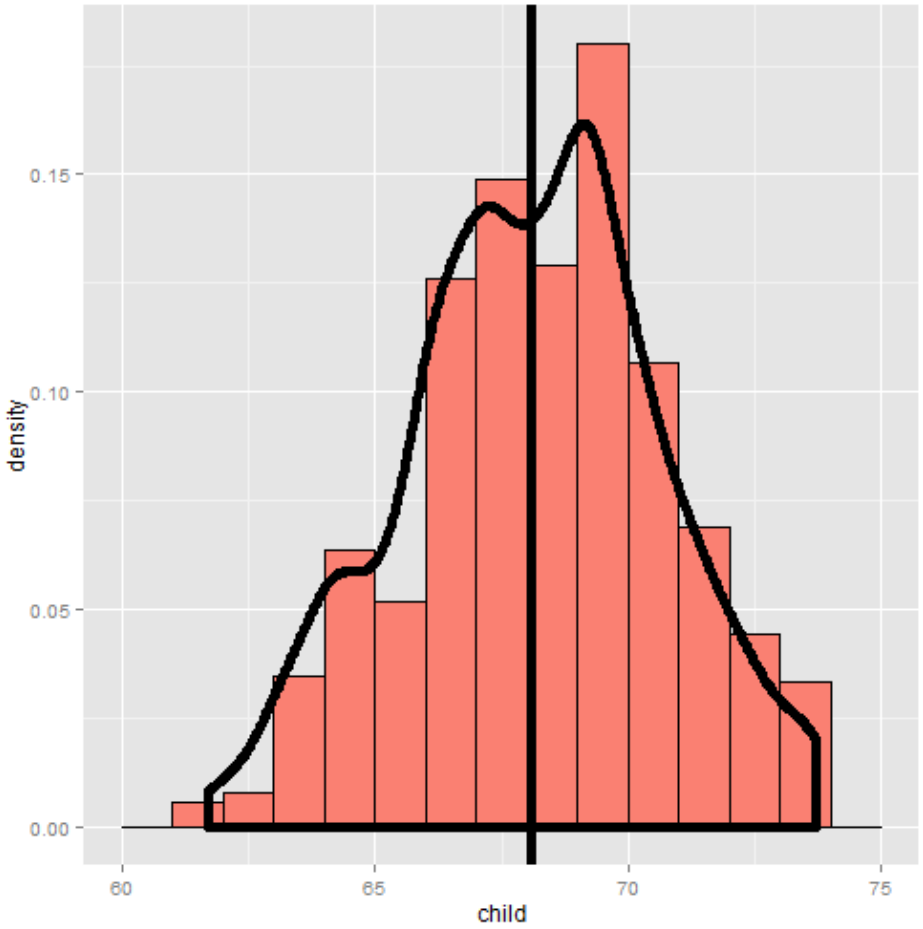
Find the center of mass of the bars



Using manipulate

```
library(manipulate)
myHist <- function(mu){
    g <- ggplot(galton, aes(x = child))
    g <- g + geom_histogram(fill = "salmon",
      binwidth=1, aes(y = ..density..), colour = "black")
    g <- g + geom_density(size = 2)
    g <- g + geom_vline(xintercept = mu, size = 2)
    mse <- round(mean((galton$child - mu)^2), 3)
    g <- g + labs(title = paste('mu = ', mu, ' MSE = ', mse))
    g
}
manipulate(myHist(mu), mu = slider(62, 74, step = 0.5))
```

The center of mass is the empirical mean

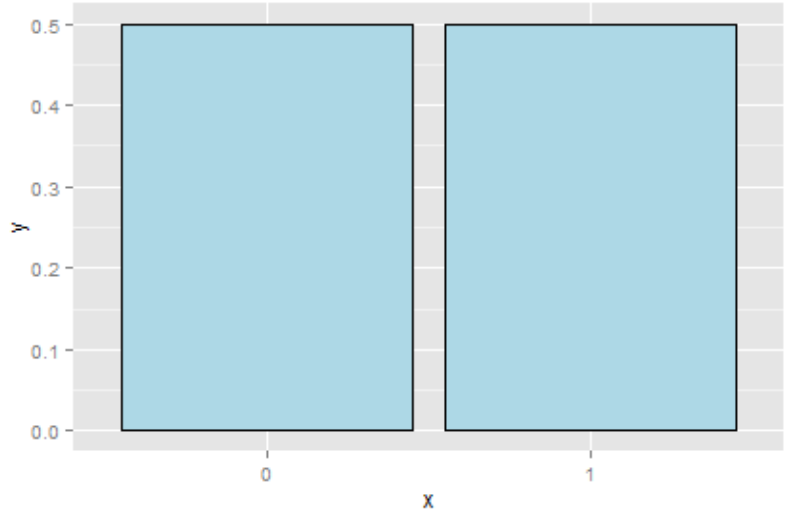


Example of a population mean

- 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 $.5$



What about a biased coin?

- Suppose that a random variable, X, is so that $P(X = 1) = p$ and $P(X = 0) = (1 - p)$
- (This is a biased coin when $p \neq 0.5$)
- What is its expected value?

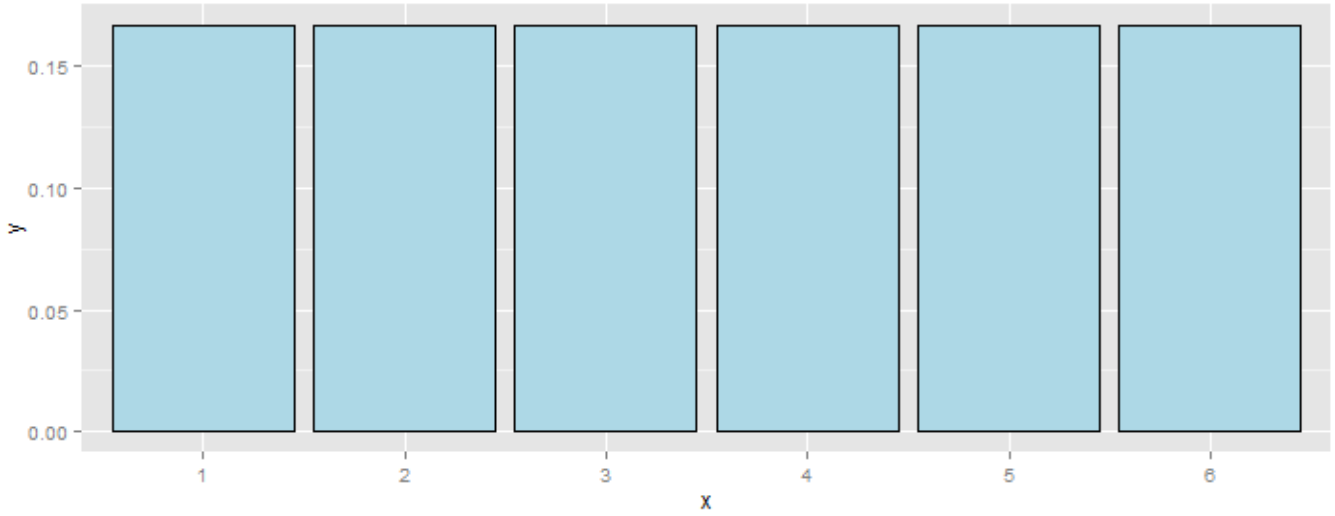
$$E[X] = 0 * (1 - p) + 1 * p = p$$

Example

- Suppose that a die is rolled and X is the number face up
- What is the expected value of X?

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

- Again, the geometric argument makes this answer obvious without calculation.

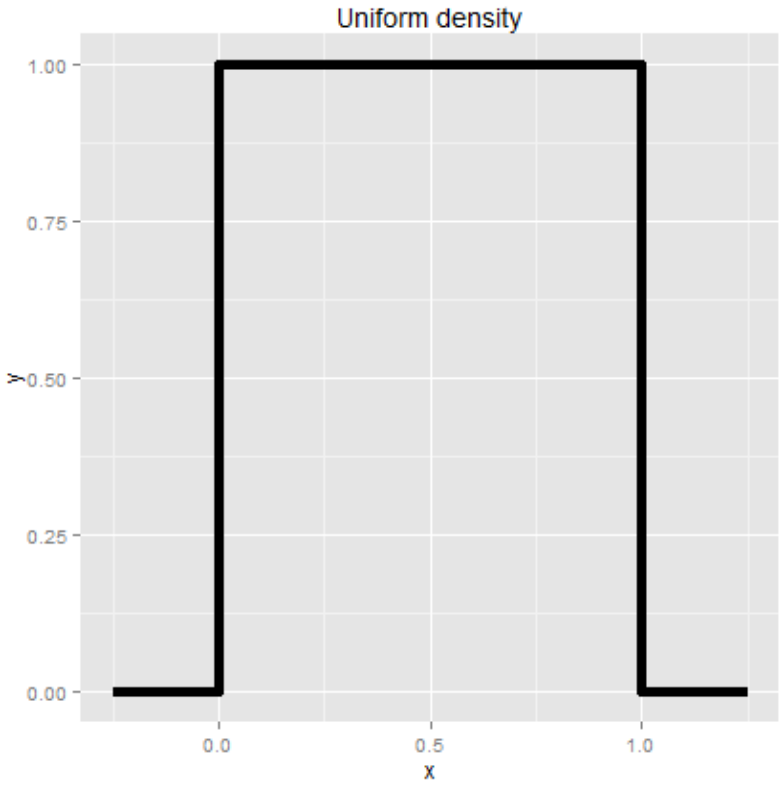


Continuous random variables

- For a continuous random variable, X, with density, f, the expected value is again exactly the center of mass of the density

Example

- Consider a density where $f(x) = 1$ for x between zero and one
- (Is this a valid density?)
- Suppose that X follows this density; what is its expected value?

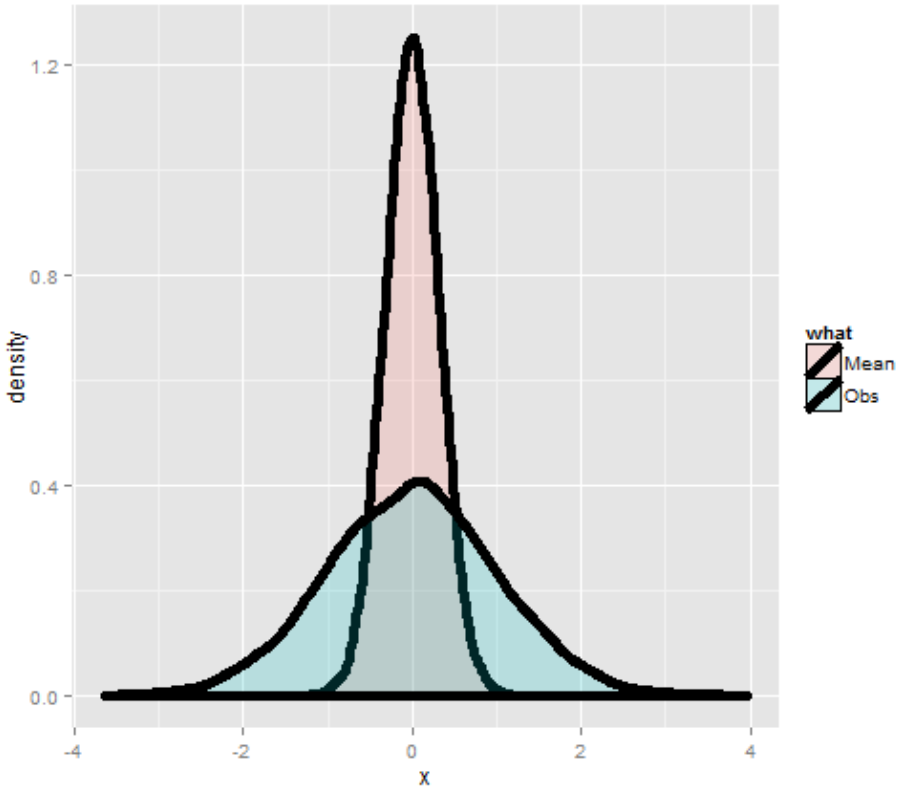


Facts about expected values

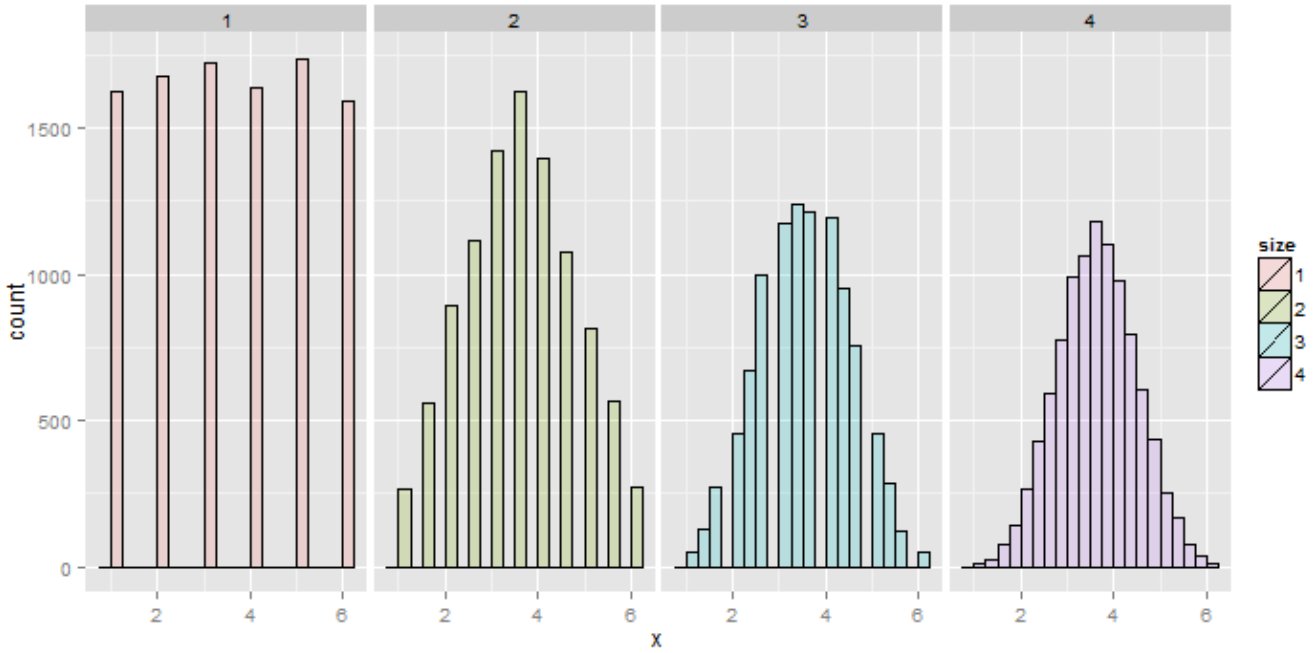
- Recall that expected values are properties of distributions
- Note the average of random variables is itself a random variable and its associated distribution has an expected value
- The center of this distribution is the same as that of the original distribution
- Therefore, the expected value of the **sample mean** is the population mean that it's trying to estimate
- When the expected value of an estimator is what its trying to estimate, we say that the estimator is **unbiased**
- Let's try a simulation experiment

Simulation experiment

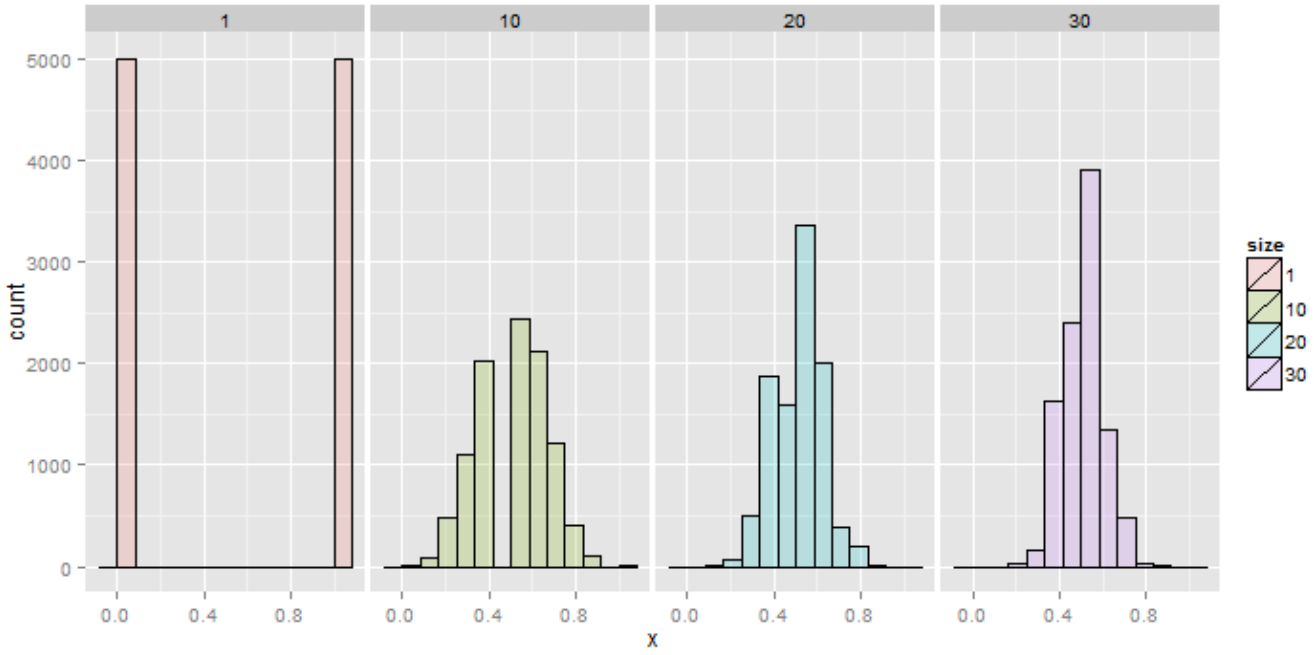
Simulating normals with mean 0 and variance 1 versus averages of 10 normals from the same population



Averages of x die rolls



Averages of x coin flips



Sumarizing what we know

- Expected values are properties of distributions
- The population mean is the center of mass of population
- The sample mean is the center of mass of the observed data
- The sample mean is an estimate of the population mean
- The sample mean is unbiased
  - The population mean of its distribution is the mean that it's trying to estimate
- The more data that goes into the sample mean, the more concentrated its density / mass function is around the population mean

Expected values

Statistical Inference

Brian Caffo, Jeff Leek, Roger Peng
Johns Hopkins Bloomberg School of Public Health

Expected values

- Expected values are useful cor characterizing a distribution
- The mean is a characterization of its center
- The variance and standard deviation are characterizations of how spread out it is
- Our sample expected values (the sample mean and variance) will estimate the population versions

The population mean

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

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

  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, $\{x, p(x)\}$

The sample mean

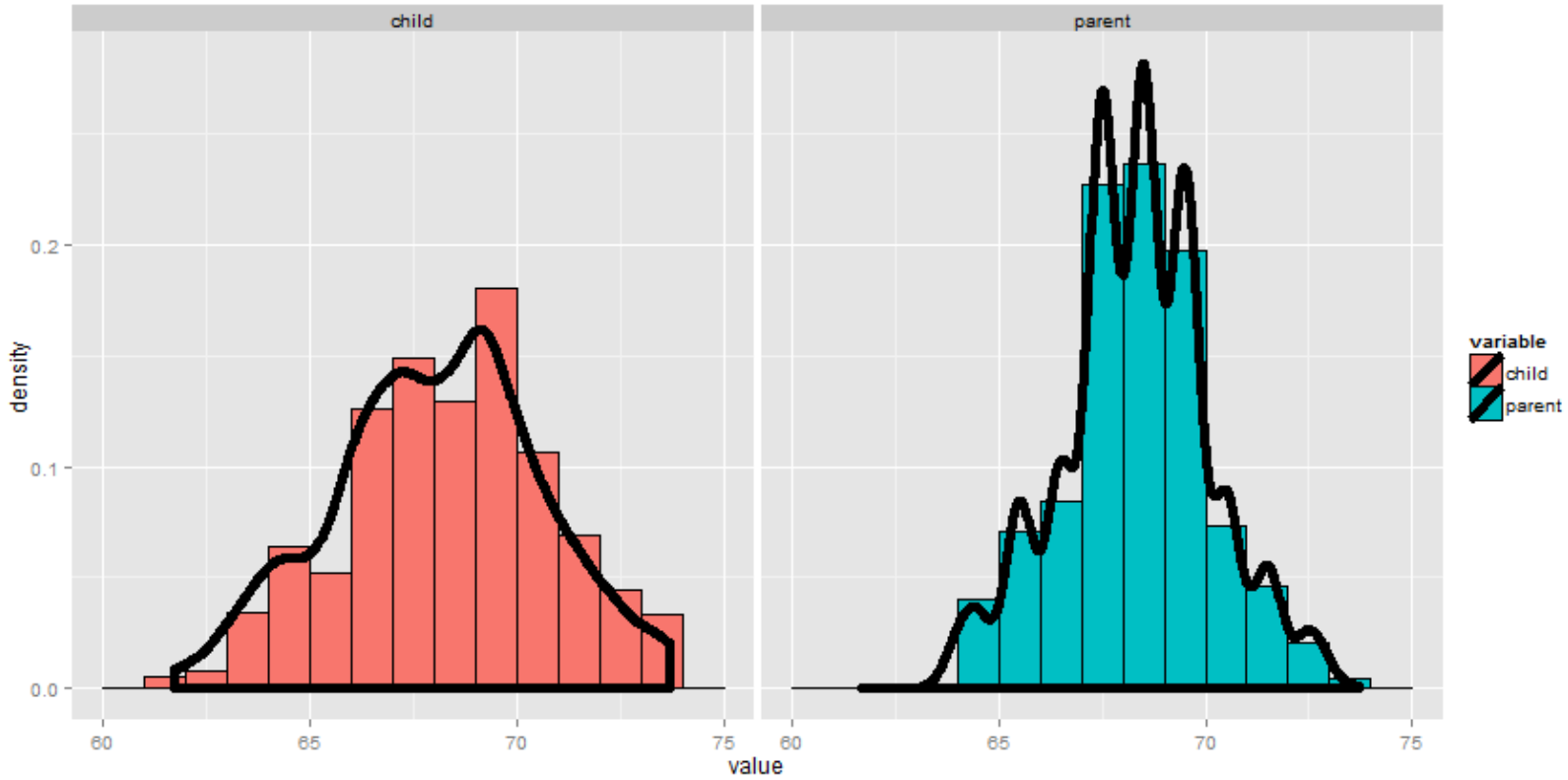
- The sample mean estimates this population mean
- The center of mass of the data is the empirical mean

$$\bar{X} = \sum_{i=1}^{n} x_i p(x_i)$$

where $p(x_i) = 1/n$

Example

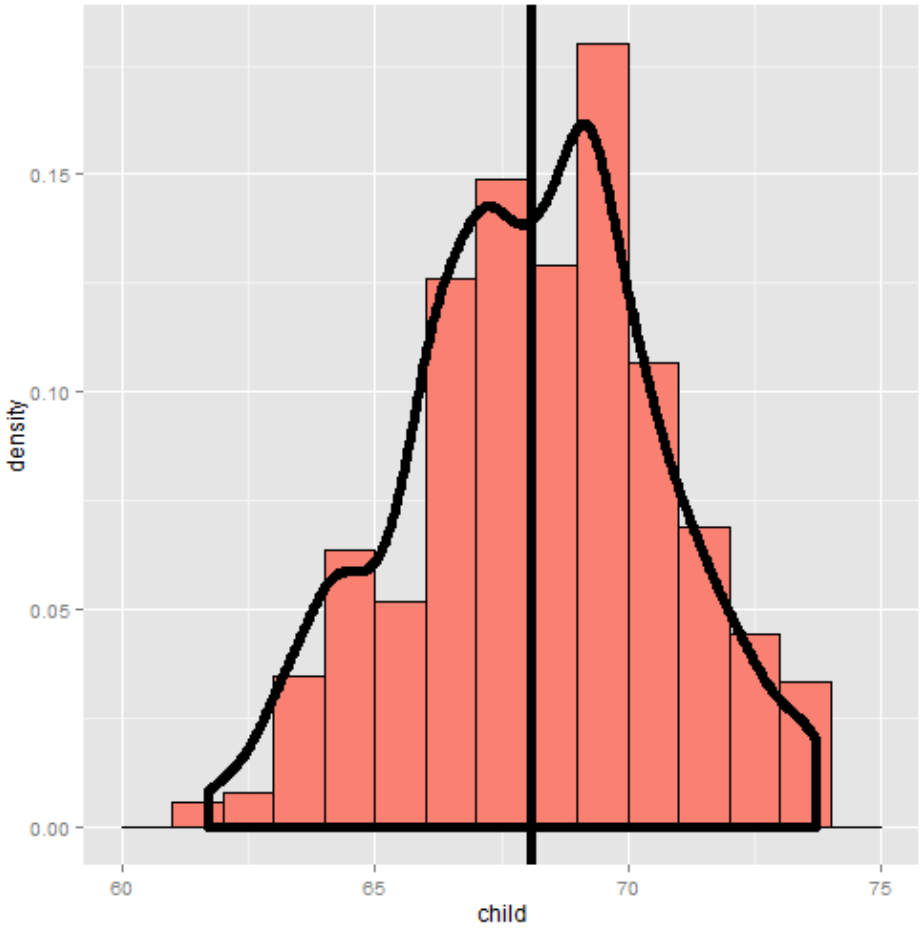
Find the center of mass of the bars



Using manipulate

```
library(manipulate)
myHist <- function(mu){
    g <- ggplot(galton, aes(x = child))
    g <- g + geom_histogram(fill = "salmon",
      binwidth=1, aes(y = ..density..), colour = "black")
    g <- g + geom_density(size = 2)
    g <- g + geom_vline(xintercept = mu, size = 2)
    mse <- round(mean((galton$child - mu)^2), 3)
    g <- g + labs(title = paste('mu = ', mu, ' MSE = ', mse))
    g
}
manipulate(myHist(mu), mu = slider(62, 74, step = 0.5))
```

The center of mass is the empirical mean

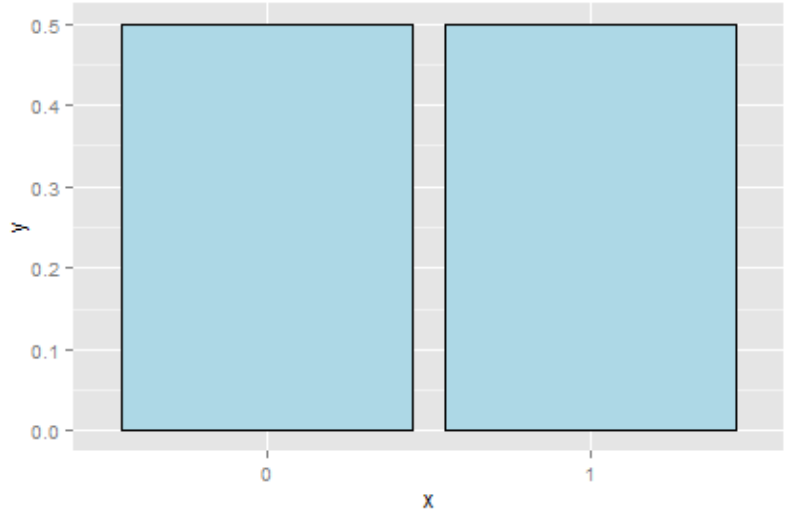


Example of a population mean

- 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 $.5$



What about a biased coin?

- Suppose that a random variable, X, is so that $P(X = 1) = p$ and $P(X = 0) = (1 - p)$
- (This is a biased coin when $p \neq 0.5$)
- What is its expected value?

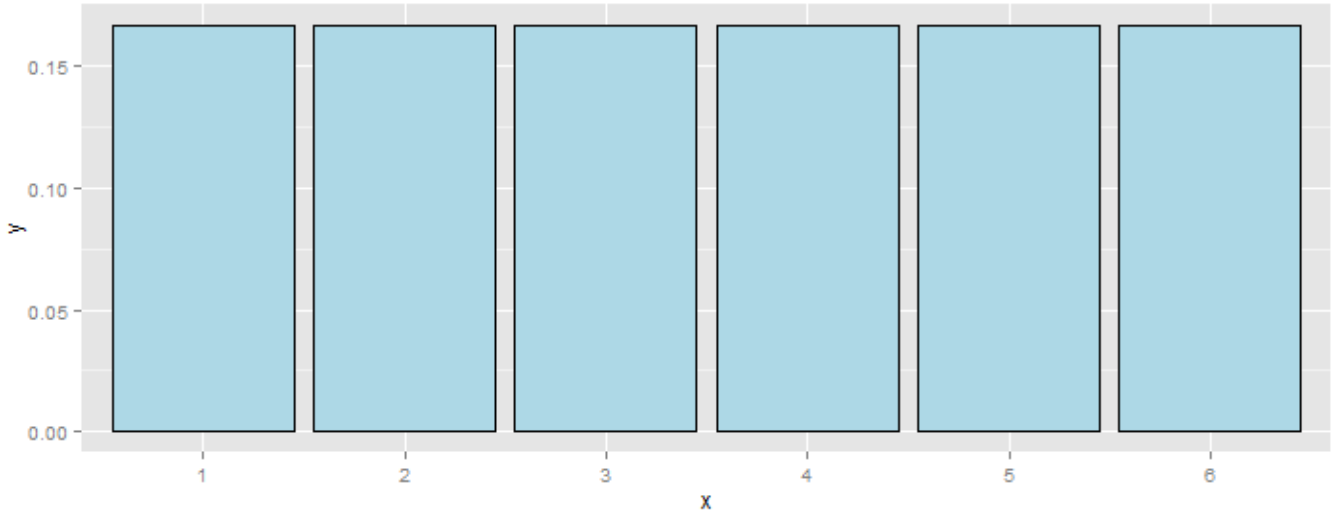
$$E[X] = 0 * (1 - p) + 1 * p = p$$

Example

- Suppose that a die is rolled and X is the number face up
- What is the expected value of X?

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

- Again, the geometric argument makes this answer obvious without calculation.

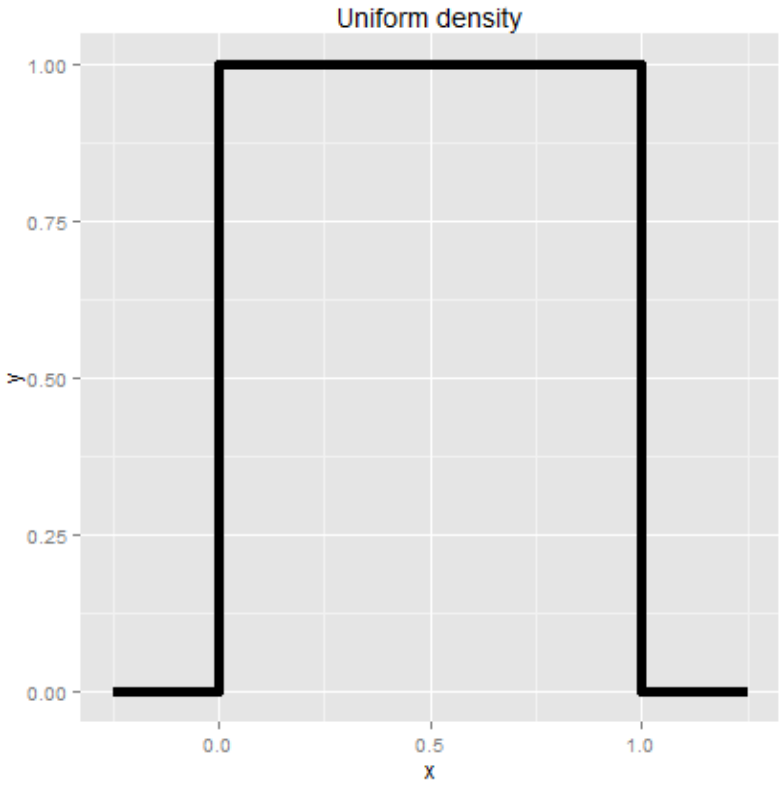


Continuous random variables

- For a continuous random variable, X, with density, f, the expected value is again exactly the center of mass of the density

Example

- Consider a density where $f(x) = 1$ for x between zero and one
- (Is this a valid density?)
- Suppose that X follows this density; what is its expected value?

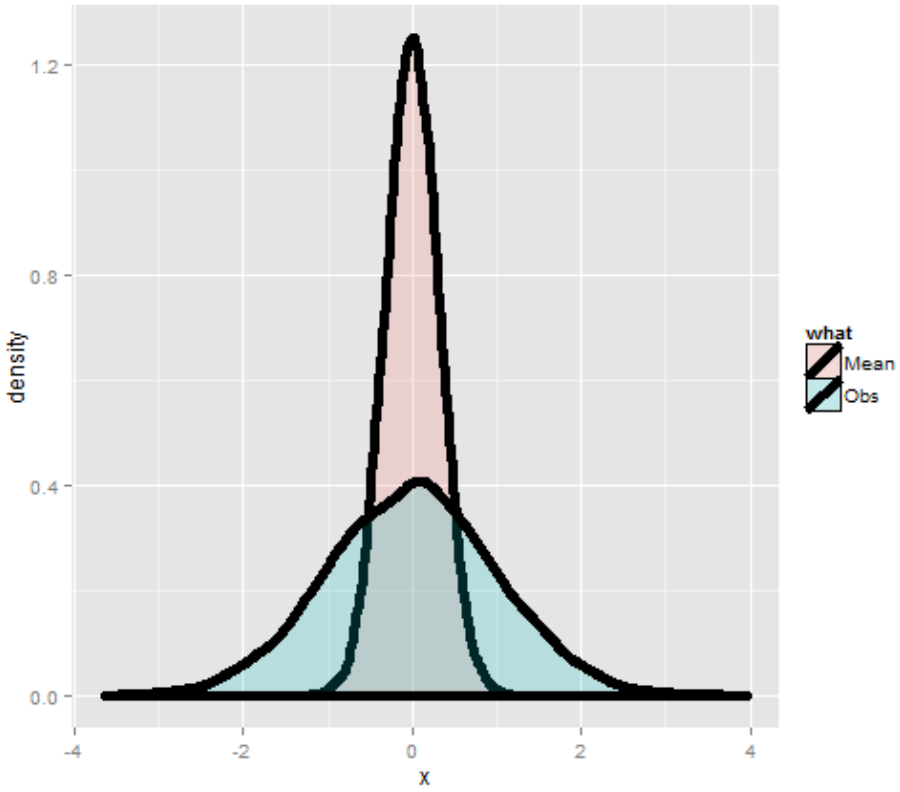


Facts about expected values

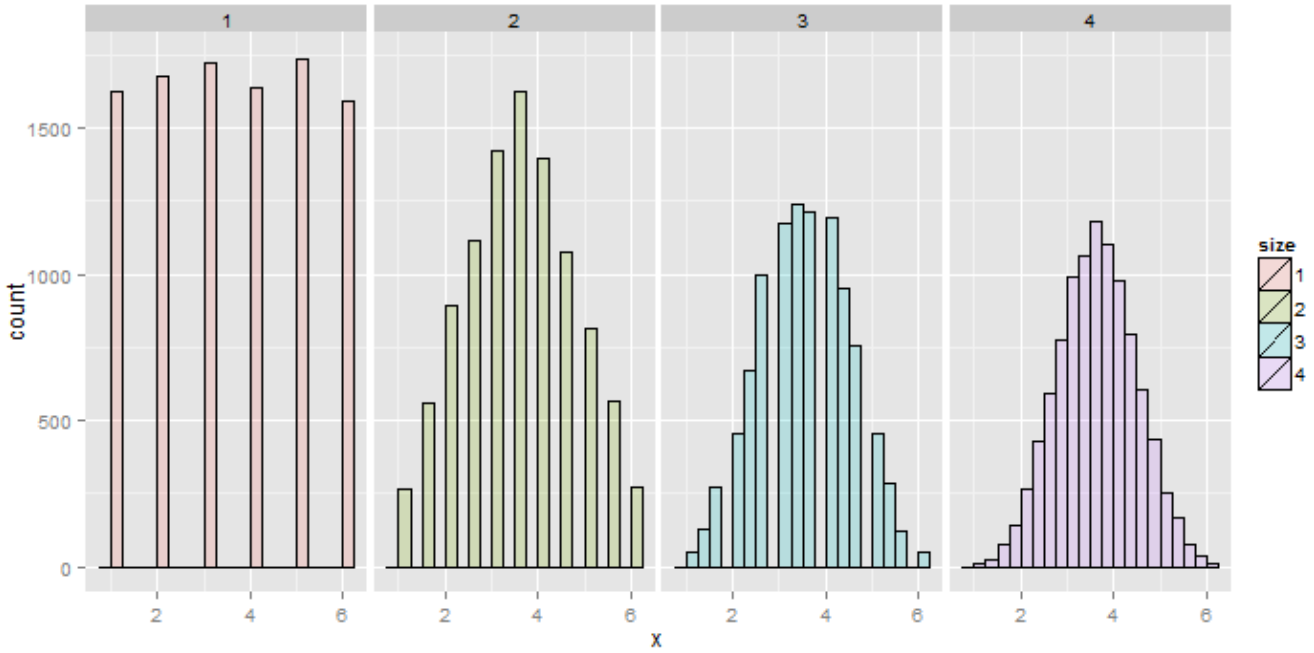
- Recall that expected values are properties of distributions
- Note the average of random variables is itself a random variable and its associated distribution has an expected value
- The center of this distribution is the same as that of the original distribution
- Therefore, the expected value of the **sample mean** is the population mean that it's trying to estimate
- When the expected value of an estimator is what its trying to estimate, we say that the estimator is **unbiased**
- Let's try a simulation experiment

Simulation experiment

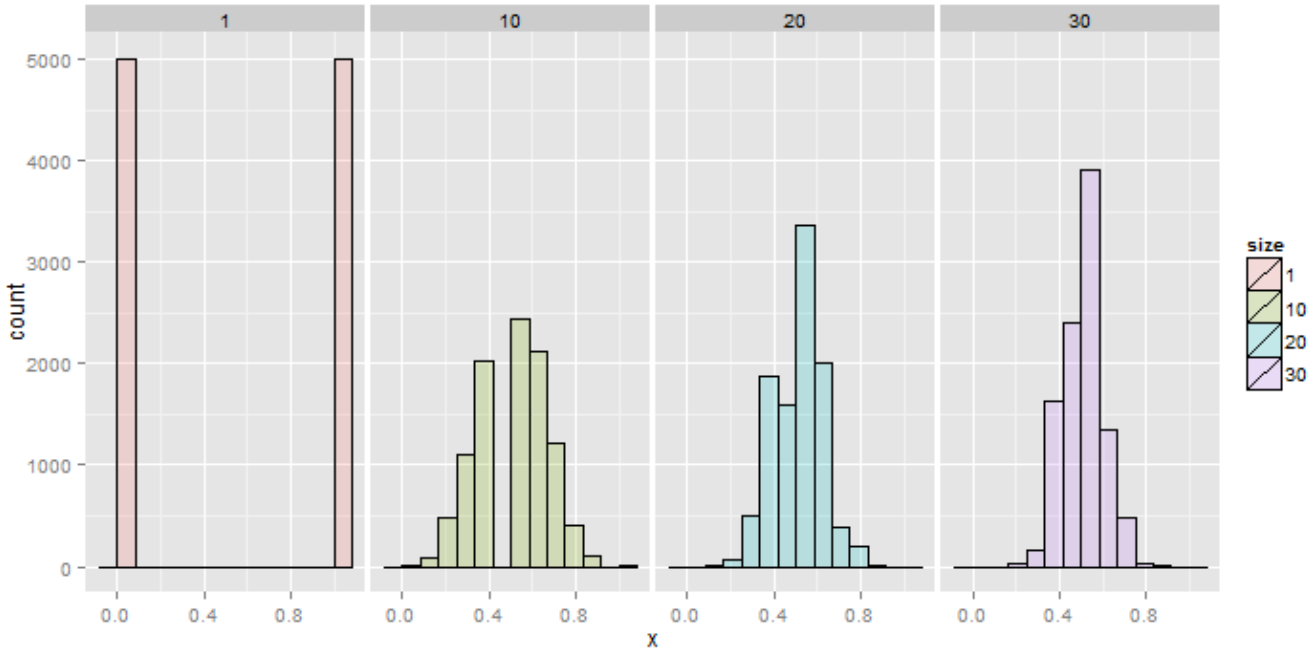
Simulating normals with mean 0 and variance 1 versus averages of 10 normals from the same population



Averages of x die rolls



Averages of x coin flips



Sumarizing what we know

- Expected values are properties of distributions
- The population mean is the center of mass of population
- The sample mean is the center of mass of the observed data
- The sample mean is an estimate of the population mean
- The sample mean is unbiased
  - The population mean of its distribution is the mean that it's trying to estimate
- The more data that goes into the sample mean, the more concentrated its density / mass function is around the population mean