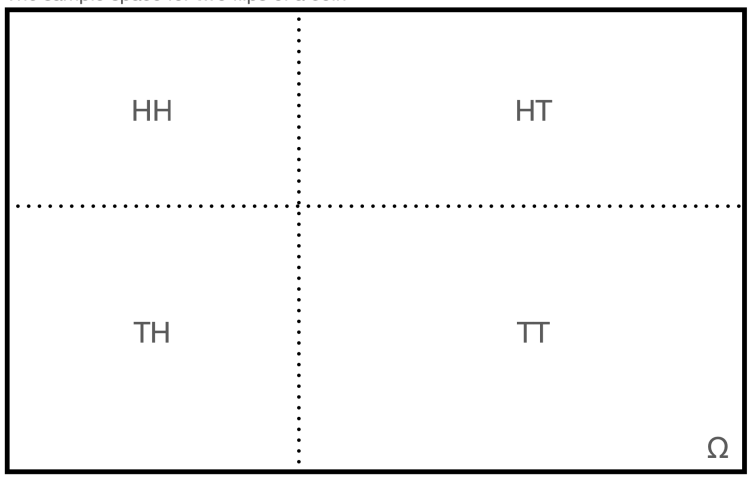
Probability Distributions and Confidence Intervals

36-600

Probability

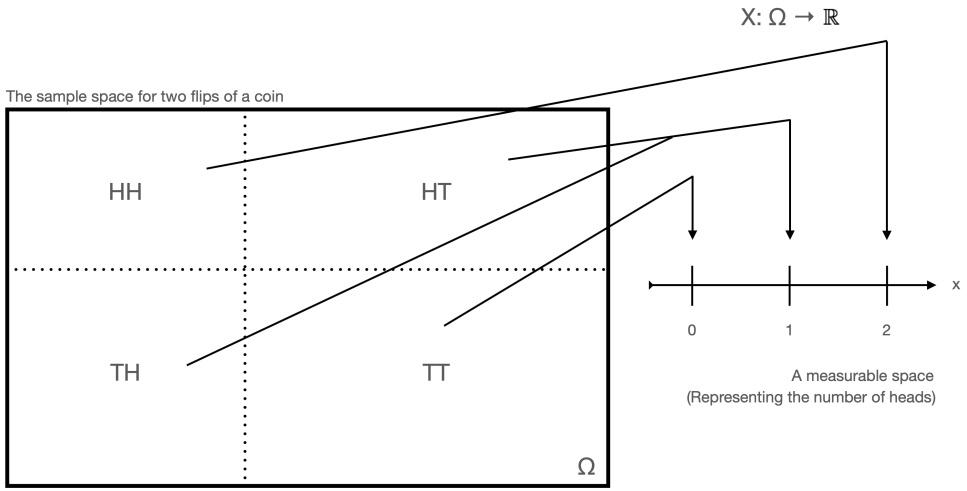
- What is probability?
  - it is, for our purposes, the long-term frequency of the occurrence of an event
- OK...what's an event?
  - an experimental outcome, or a set of experimental outcomes
  - examples: observing heads then tails when flipping a coin twice, or observing at least one head when flipping a coin twice, etc.
- How do we organize the information about experimental outcomes?
  - for a given experiment, all possible outcomes (or *simple events*) comprise a *sample space*, conventionally denoted Ω

The sample space for two flips of a coin



Random Variables

- Working directly with sample spaces is tedious: instead, we work with *random variables*
- A random variable is a measurable function $X : \Omega \to S$ mapping the sample space Ω to a measurable space
  - intuition: there is no "measurable distance" between `HH` and `TT` in Ω, but the distance is 2 along the real-number line



- Note that we generally view random variables not as functions, but as function outputs (i.e., data samples), which may be...
  - continuously valued (perhaps representing time, or distance, etc.); or
  - discretely valued (perhaps representing counts, or choices from a finite menu, etc.)

Probability Distributions

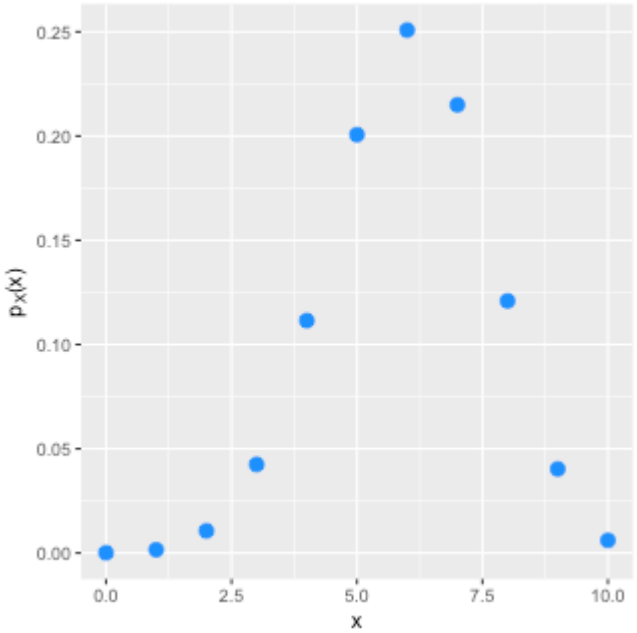
- Neither a sample space, nor a random variable defined given that sample space, have any implicit notion of event probability
- A *probability distribution* $P : \Omega \to \mathbb{R}$ is what describes how probabilities are distributed across the output values of a random variable

- For instance, we can specify that the probability of seeing x heads in k coin flips is given by the *binomial distribution* (or specifically its probability mass function):

$$p_X(x) = \binom{k}{x} p^x (1-p)^{k-x} = \frac{n!}{x!(k-x)!} p^x (1-p)^{k-x},$$

  where $x \in \{0, 1, \ldots, k\}$, k is the number of *binomial trials*, and $x! = x \cdot (x-1) \cdot (x-2) \cdots 2 \cdot 1$ is the factorial function

- There are many commonly used continuous and discrete distributions, including the normal (or Gaussian), gamma (and exponential), beta, uniform, and Poisson distributions



Probability Distributions: Motivation

- You may be asking right now: "why exactly are we learning about probability distributions?"
- Because they underlie
  - the sampling of random variables, i.e., the data you wish to analyze;
  - the generation of confidence intervals and performance of hypothesis tests; and
  - the details of statistical learning models, particularly ones like linear regression and Naive Bayes
- The following picture illustrates the statistical inference "cycle," which relies on the sampling of random variables (i.e., data) from distributions (i.e., populations):

OK...But What is a Statistic?

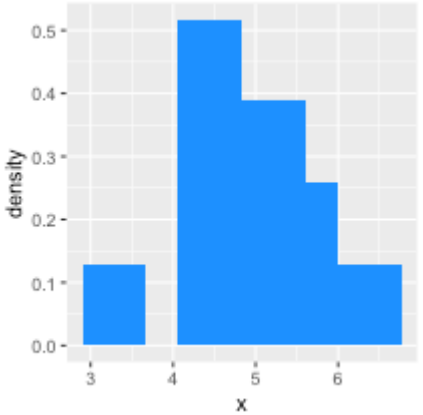
- A *statistic* is simply a function of observed data
  - it "summarizes" an array of numbers into a single number
  - because it is a function of random variables, a statistic is itself a random variable sampled from some distribution, dubbed a *sampling distribution*
- If we assume that the data we collect are *independent and identically distributed*, or *iid*, samples from some distribution, we can...
  - try to derive the sampling distribution of our statistic, then
  - use that sampling distribution and the observed statistic value to construct a confidence interval for, e.g., the distribution mean, or, alternatively,
  - use that sampling distribution and the observed statistic value to test a preconceived notion (or *null hypothesis*) about, e.g., the distribution mean
- What if we cannot assume a distribution for our sampled data?
  - we can fall back upon bootstrapping (more on that later)
- What if we cannot determine the sampling distribution for our statistic?
  - we can fall back upon simulations (which are beyond the current scope of this course)

Example

- Let's assume that we run an experiment and that we observe the values

```
##  [1] 4.673964 5.552462 4.325056 5.214359 5.310769 6.173966 5.618790 4.887266
##  [9] 5.917028 4.776741 5.526448 4.205156 6.427756 3.533180 4.763317 4.806662
## [17] 4.150245 5.058465 4.182330 2.949692
```

- The data are sampled from some distribution with some mean and some width...we can use methods of statistical inference to try to say something about each
- Perhaps, here, we will assume that the data are normally distributed (i.e., like a bell curve) and we want to be able to infer the (unknown!) true mean



- What statistic might we use to summarize these data and to try to infer the mean?
  - perhaps the *sample mean* $\bar{X}$ or the *sample median* (the middle sorted value)
  - in general, the sample mean is best for inference

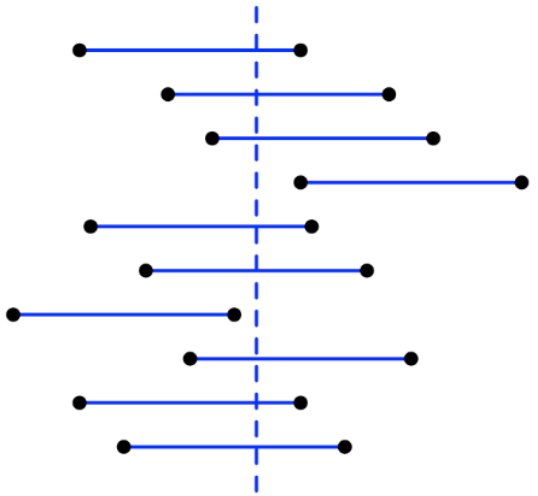
Confidence Intervals

- A *confidence interval* is a random interval $[\hat{\theta}_L, \hat{\theta}_U]$ that satisfies the condition

$$P(\hat{\theta}_L \leq \theta \leq \hat{\theta}_U) = 1 - \alpha\,,$$

  where θ is a population parameter of interest (like the mean) and $1 - \alpha$ is the confidence coefficient (which, e.g., has value 0.95 when we are constructing a two-sided 95% confidence interval)

- Every time we run an experiment, we generate new data, meaning we generate new statistic values and new intervals
  - so how do we interpret any one interval?
  - we say that any one evaluated interval has a $100(1 - \alpha)$-percent chance of *overlapping the true value* of θ
  - we do *not* say that the probability that θ lies within the evaluated interval is $1 - \alpha$



Confidence Intervals

- There is much math that goes into the construction of confidence intervals and many formulas for computing them that one would see in, e.g., introductory statistics classes
  - for instance, if we sample n data from a normal distribution with an unknown mean μ and a *known* standard deviation σ, and our sample mean is $\bar{X}$, then a two-sided 95% confidence interval is

$$\bar{X} \pm 1.96 \frac{\sigma}{\sqrt{n}}$$

- However, a more straightforward, more general, and more math-free approach is to numerically construct two-sided confidence intervals for a population parameter θ using codes like this:

```
f <- function(theta,[arguments])
{
  [cumulative distribution function]([arguments]) - q
}
uniroot(f,interval=[interval],[arguments],q=1-alpha/2)$root
uniroot(f,interval=[interval],[arguments],q=alpha/2)$root
```

- Note that a *cumulative distribution function* takes a coordinate x as input, and outputs the probability that we would observe a data sample with value $\leq x$

Example

- Let's assume that we ran an experiment and that we observed the values

```
##  [1] 4.673964 5.552462 4.325056 5.214359 5.310769 6.173966 5.618790 4.887266
##  [9] 5.917028 4.776741 5.526448 4.205156 6.427756 3.533180 4.763317 4.806662
## [17] 4.150245 5.058465 4.182330 2.949692
```

- What is a two-sided 95% confidence interval for the population mean?

- Setting up the solution requires knowing some details about distributions: in this course, we will provide those details
- If we assume that each of the $n = 20$ data are sampled from a normal distribution with some mean and some standard deviation (i.e., some "width"), then

$$T = \frac{\bar{X} - \mu}{S/\sqrt{n}} \sim t_{n-1} \, ,$$

  i.e., the stated combination of the sample mean $\bar{X}$ and sample standard deviation S is sampled from a t-distribution for $n - 1$ degrees of freedom

Example

- Given this information, here's the code:

```
alpha <- 0.05
f <- function(mu,x.bar,s,n,q)
{
  pt((x.bar-mu)/(s/sqrt(n)),n-1) - q # "p" means cumulative distribution function
                                     # "t" means t-distribution
}
uniroot(f,interval=c(-100,100),x.bar=mean(x),s=sd(x),n=length(x),q=1-alpha/2)$root
```

```
## [1] 4.49701
```

```
uniroot(f,interval=c(-100,100),x.bar=mean(x),s=sd(x),n=length(x),q=alpha/2)$root
```

```
## [1] 5.308358
```

- There is a 95% chance that the random interval $[4.497, 5.308]$ overlaps the true distribution mean
  - note: this is *not* the same as saying that the probability that the mean lies in the stated interval is 0.95!
- We would derive the same interval if we were to use the `R` function `t.test()`
- For more details about confidence intervals, see, e.g., Chapter 1 of this book

Bootstrap Confidence Intervals

- Let's assume we have the same data as in the example, but that we will not (or cannot) make the assumption that the data are normally distributed...and in fact, we cannot assume that the data follow *any* known distribution

  $\Rightarrow$ to construct a confidence interval for the mean, we can utilize *bootstrapping*

- The bootstrapping algorithm is simple!
  - generate k new datasets of size n from the observed data by sampling *with replacement*
  - for each of the k datasets, generate the sample mean
  - determine, e.g., the 2.5^{th} and 97.5^{th} percentiles of the empirical sample mean distribution: that's the bootstrap confidence interval
- What is sampling "with replacement"? It means that our new dataset can have some repeated rows, and some rows that are left out

```
set.seed(101)
rows <- 1:8  # we have an original dataset with 8 rows
sort(sample(rows,length(rows),replace=TRUE)) # rows 1 and 7 appear three times, etc.
```

```
## [1] 1 1 1 2 6 7 7 7
```

- Why do we not always compute bootstrap confidence intervals?
  - because, in general, the computed intervals are too short

Example

- Let's assume that we ran an experiment and that we observed the values

```
##  [1] 4.673964 5.552462 4.325056 5.214359 5.310769 6.173966 5.618790 4.887266
##  [9] 5.917028 4.776741 5.526448 4.205156 6.427756 3.533180 4.763317 4.806662
## [17] 4.150245 5.058465 4.182330 2.949692
```

- What is a two-sided 95% bootstrap confidence interval for the population mean?

```
set.seed(101)
k     <- 10000
x.bar <- rep(NA,k)
for ( ii in 1:k ) {
  s         <- sample(length(x),length(x),replace=TRUE)
  x.bar[ii] <- mean(x[s])
}
quantile(x.bar,probs=c(0.025,0.975))
```

```
##     2.5%    97.5%
## 4.526950 5.261799
```

- Compare this result to $[4.497, 5.308]$, the interval derived assuming normality