A/B Testing and Beyond

Designed Experiments for Data Scientists

Week 1

Wednesday September 13th, 2017

DATA INSTITUTE
UNIVERSITY OF SAN FRANCISCO

Outline

- Recap
- Instagram Experiment
  - Discussion
- Statistical Prerequisites
  - Random Variables and Distributions
  - Point and Interval Estimation
  - Hypothesis Testing

DATA INSTITUTE
UNIVERSITY OF SAN FRANCISCO

RECAP

Recap

- Response variables
- Factors (3 types) – Levels
- Experimental Conditions
- Experimental Units
- Experiments vs. Observational Studies
- QPDAC
- Randomization
- Replication
- Blocking

INSTAGRAM EXPERIMENT

Instagram Experiment

While sponsored ads serve as a source of revenue for Instagram, they also serve as a source of frustration and annoyance to users. Thus, we would like to run an experiment to gain insight into the interplay between ad revenue, user engagement and factors such as ad frequency, ad type (photo/video), whether the ad's content is targeted or not, etc. Ultimately the goal is to identify a condition that maximizes ad revenue without simultaneously plummeting user engagement below some minimally acceptable threshold.

How would you design such an experiment?

STATISTICAL PREQUISITES

Random Variables & Distributions

A random variable is a function that assigns real numbers to the outcomes of a random process.

$$Y: \Omega \rightarrow \mathbb{R}$$

We typically denote random variables with upper case letters, such as Y, and the values they can take on with lower case letters, such as y.

We refer to the possible values a random variable can take on as its support set.

Random Variables & Distributions

- A discrete random variable Y is a random variable that can only take on a finite or countably infinite number of values.
  - Example: $y = 0,1,2, \dots, n$ or $y = 0,1,2, \dots$
  - We typically associate discrete random variables with counting things

- A continuous random variable Y is a random variable that can take on any value in any subinterval of the real numbers.
  - Example: $y \geq 0$ or $y \in [0,1]$ or $-\infty < y < \infty$
  - We typically associate continuous random variables with measuring things

DATA INSTITUTE
UNIVERSITY OF SAN FRANCISCO

Random Variables & Distributions

Example 1: Suppose we send an email survey to $n = 30$ individuals and we're interested in the the number of these individuals that respond to the survey. Let Y represent the number of survey responses. In this case the support set is $y = 0,1,2, \ldots, 100$, and so Y is a discrete random variable.

Random Variables & Distributions

Example 2: Interest often lies in measuring lifetimes of people, products, and processes. Suppose that, in particular, we are interested in the lifetime of an iPhone's battery. Let Y represent the lifetime (in hours) of an iPhone battery. In this case the support set is theoretically $y \geq 0$, which is a continuous subinterval of the real numbers, and so Y is a continuous random variable.

Random Variables & Distributions

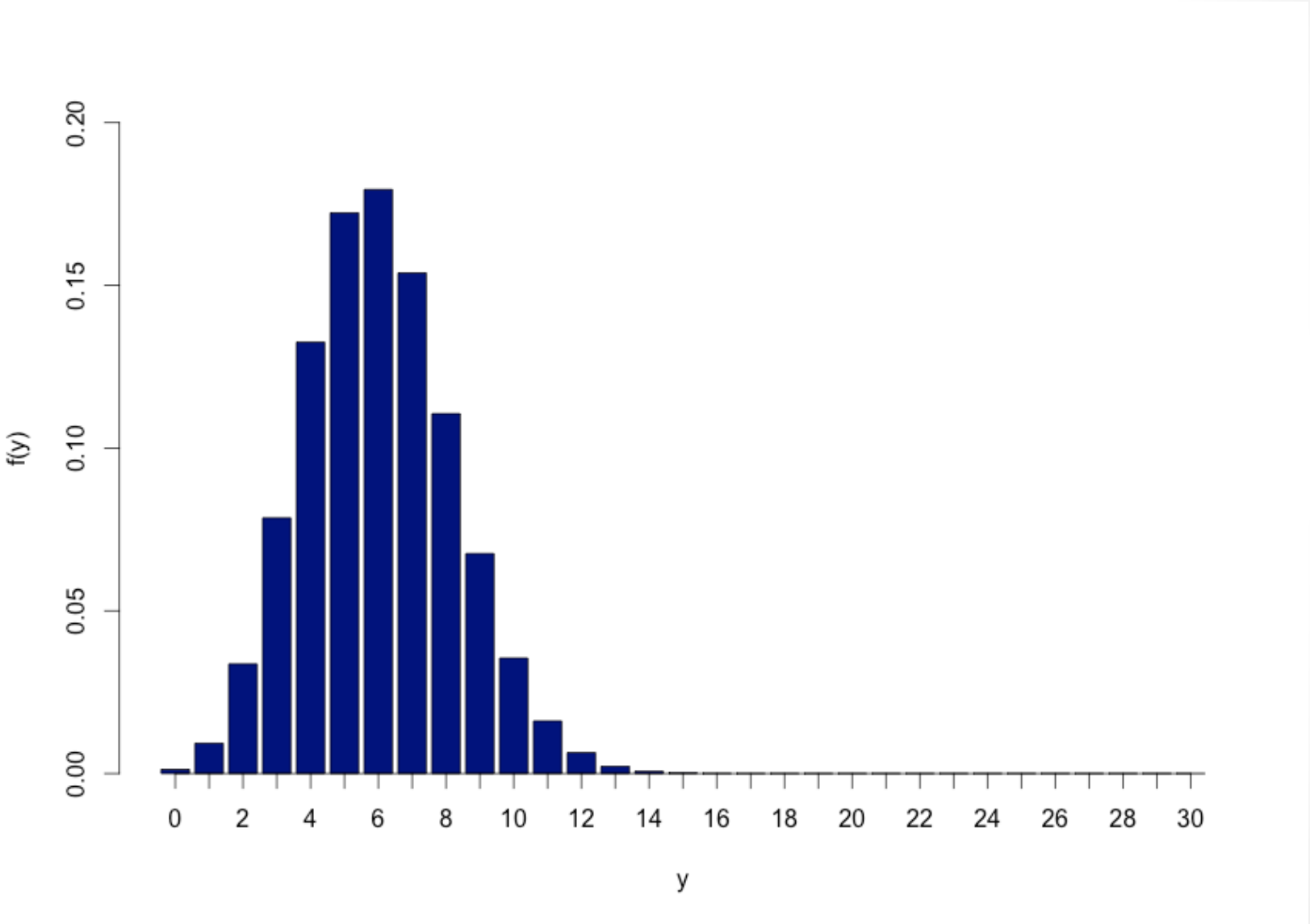
Because Y takes on values randomly, interest lies in quantifying the probability that Y assumes a particular value or lies in some interval

$$P(Y = a) = ? \quad \text{or} \quad P(a < Y < b) = ?$$

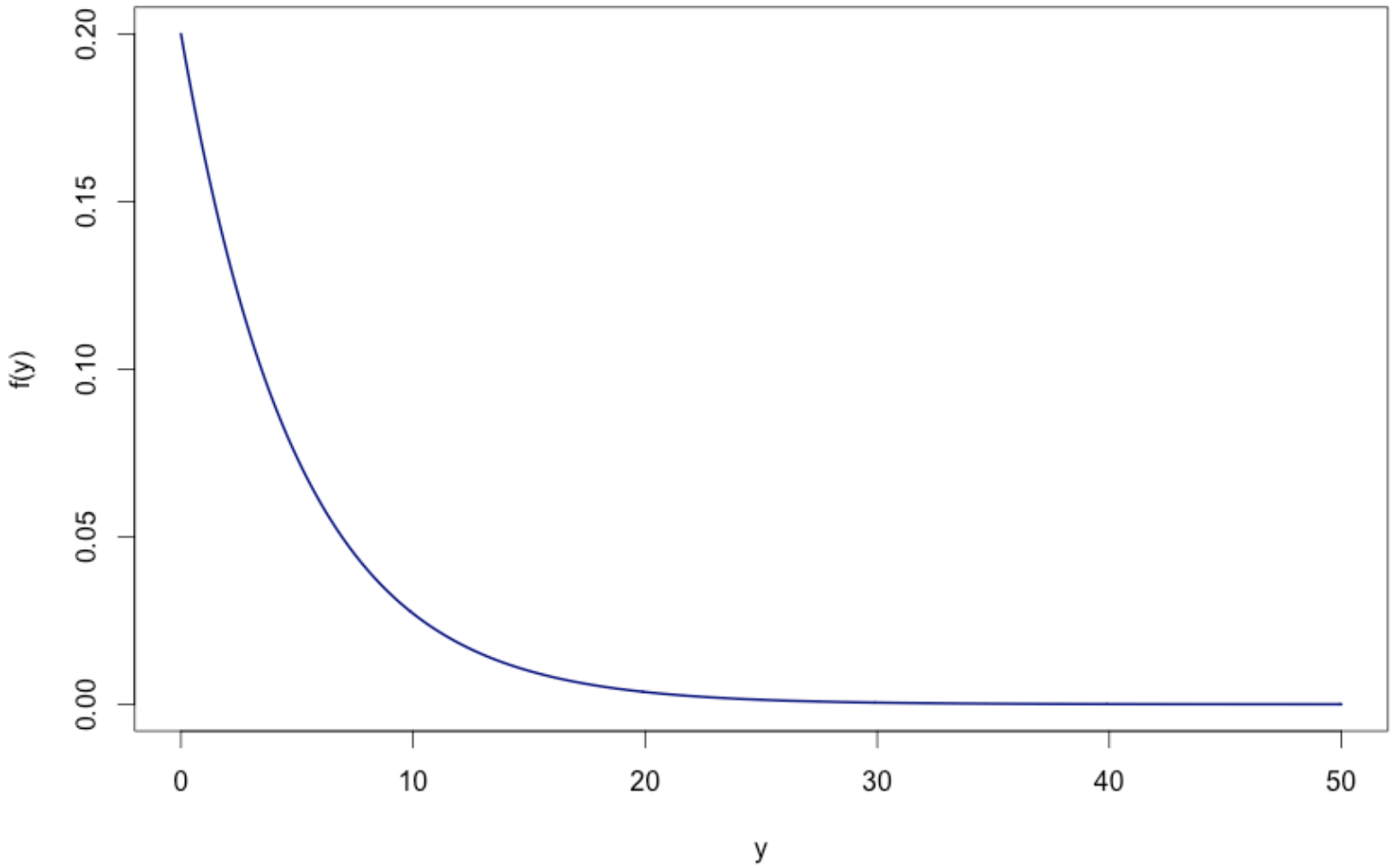
In general, Questions like these are answered with probability functions $f(y)$. The manner in which they are used depends on whether Y is discrete or continuous.

But in both cases a plot of $f(y)$ vs. y is useful for visualizing the probability distribution

Random Variables & Distributions



Random Variables & Distributions



Random Variables & Distributions

A probability mass function (PMF) is the name given to the probability function for a discrete random variable Y, and it is defined as

$$f(y) = P(Y = y)$$

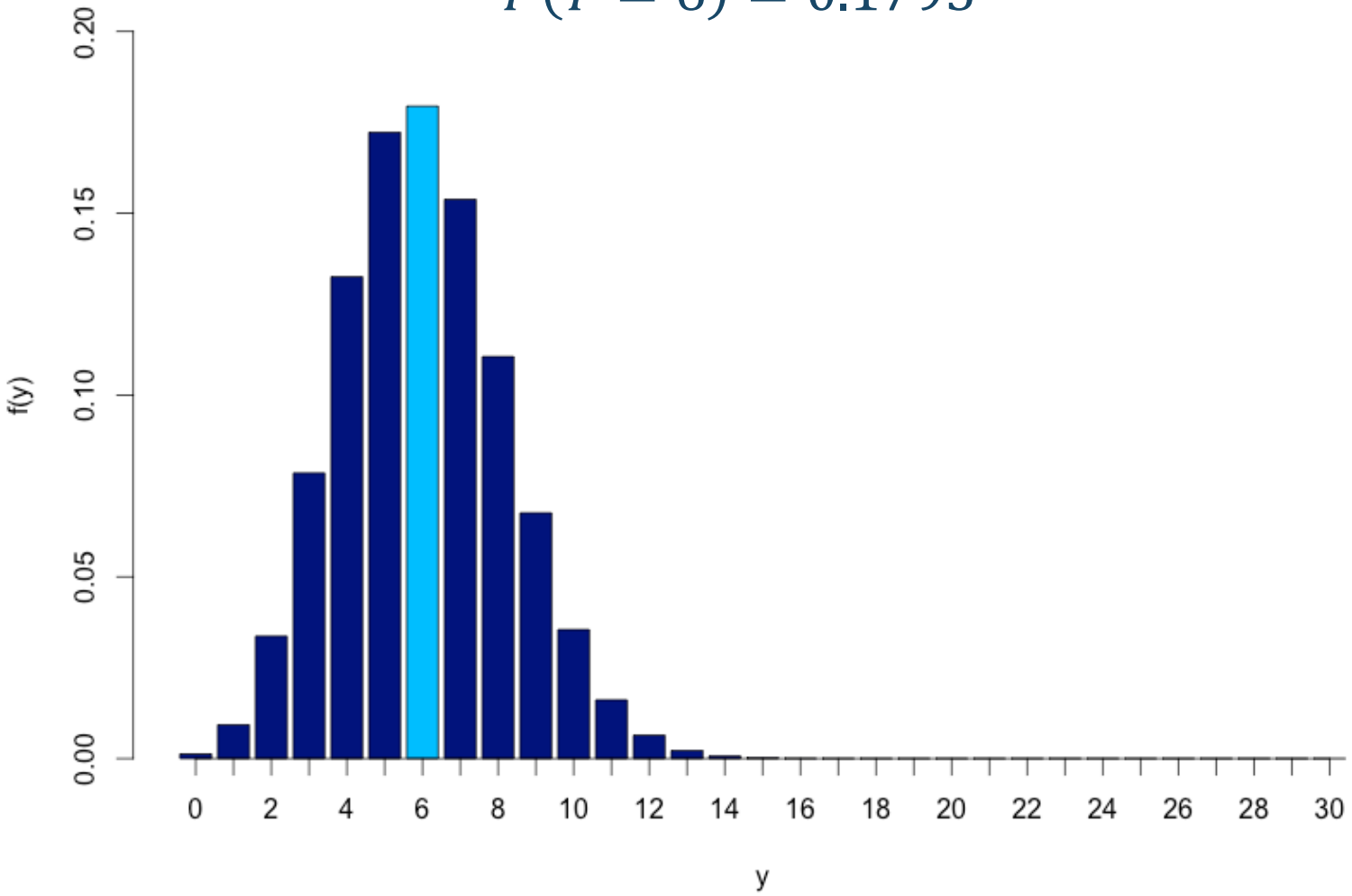
for $y \in A$, where the support set A depends on the context of the problem.

- The PMF describes the probabilistic behavior of Y;
- It allocates probability to every element of the support set.
- The PMF provides a way to characterize the probability distribution of Y.

DATA INSTITUTE
UNIVERSITY OF SAN FRANCISCO

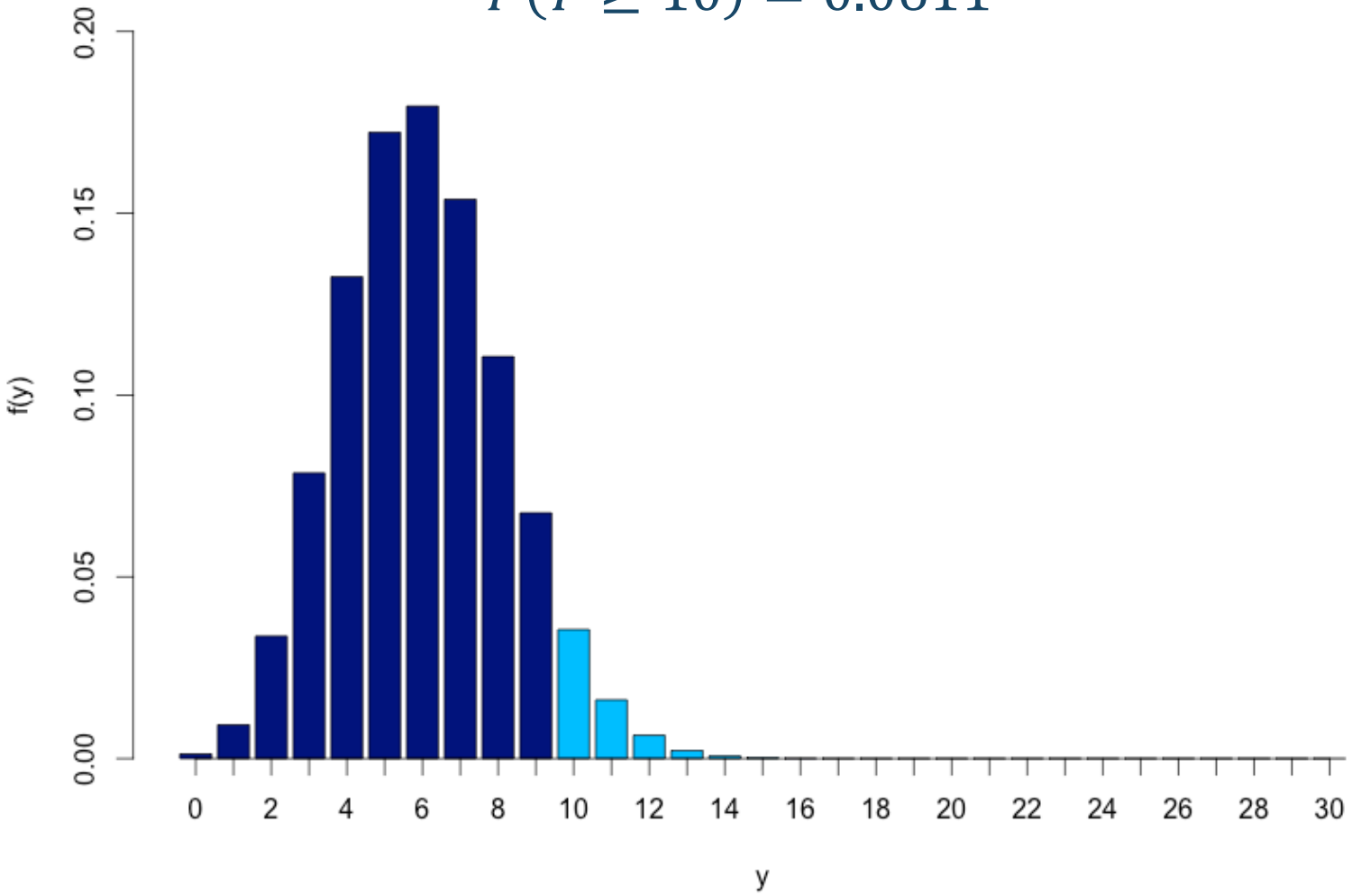
Random Variables & Distributions

$$P(Y = 6) = 0.1795$$



Random Variables & Distributions

$$P(Y \geq 10) = 0.0611$$



Random Variables & Distributions

A probability density function (PDF) is the name given to the probability function for a continuous random variable Y, and it is also denoted by

$$f(y)$$

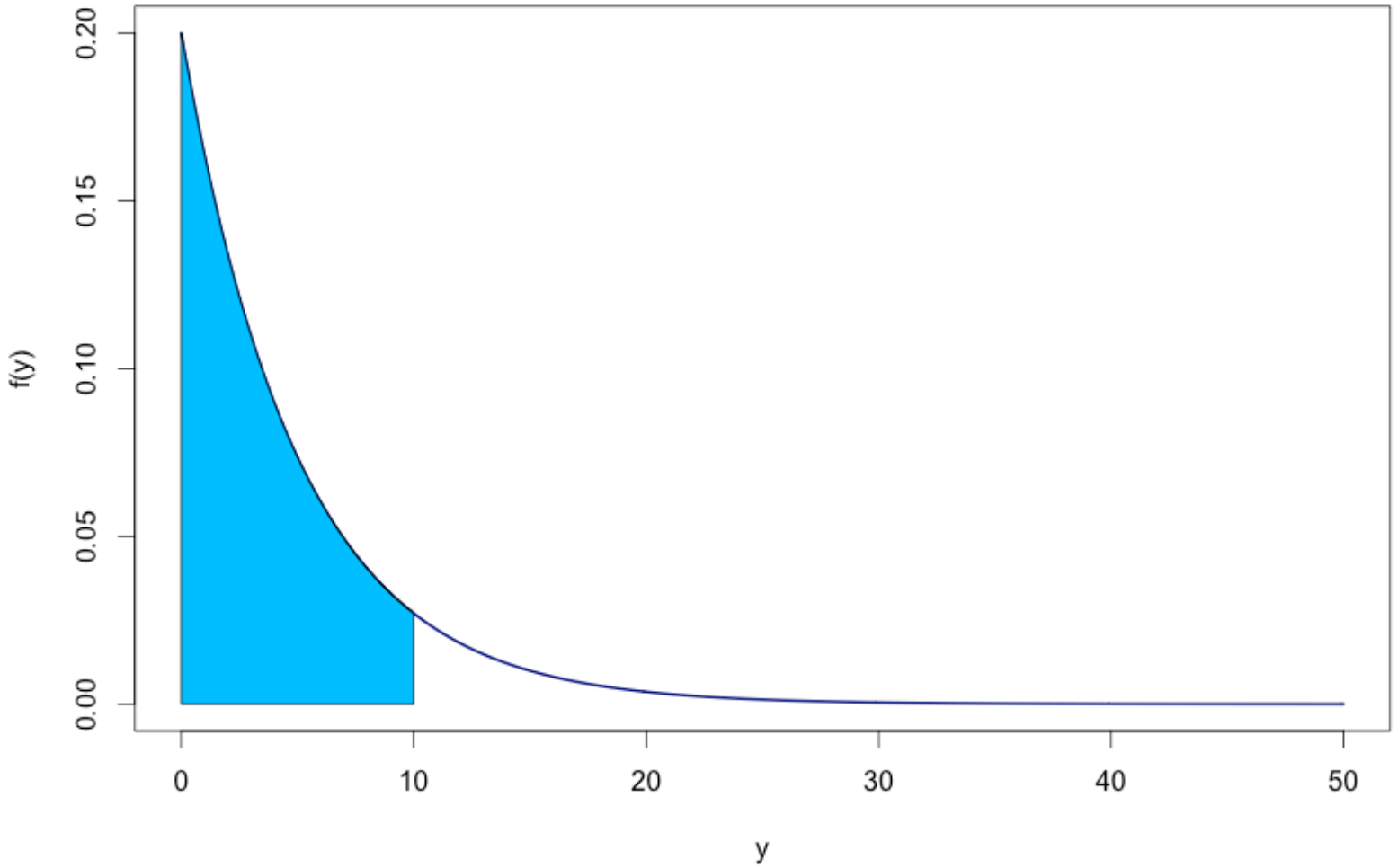
and defined for $y \in A$, where the support set A depends on the context of the problem.

- The difference between a PDF and a PMF lies in how it is used to calculate probabilities.
- If Y is a continuous random variable, we calculate probabilities associated with it by calculating areas beneath the associated PDF.

DATA INSTITUTE
UNIVERSITY OF SAN FRANCISCO

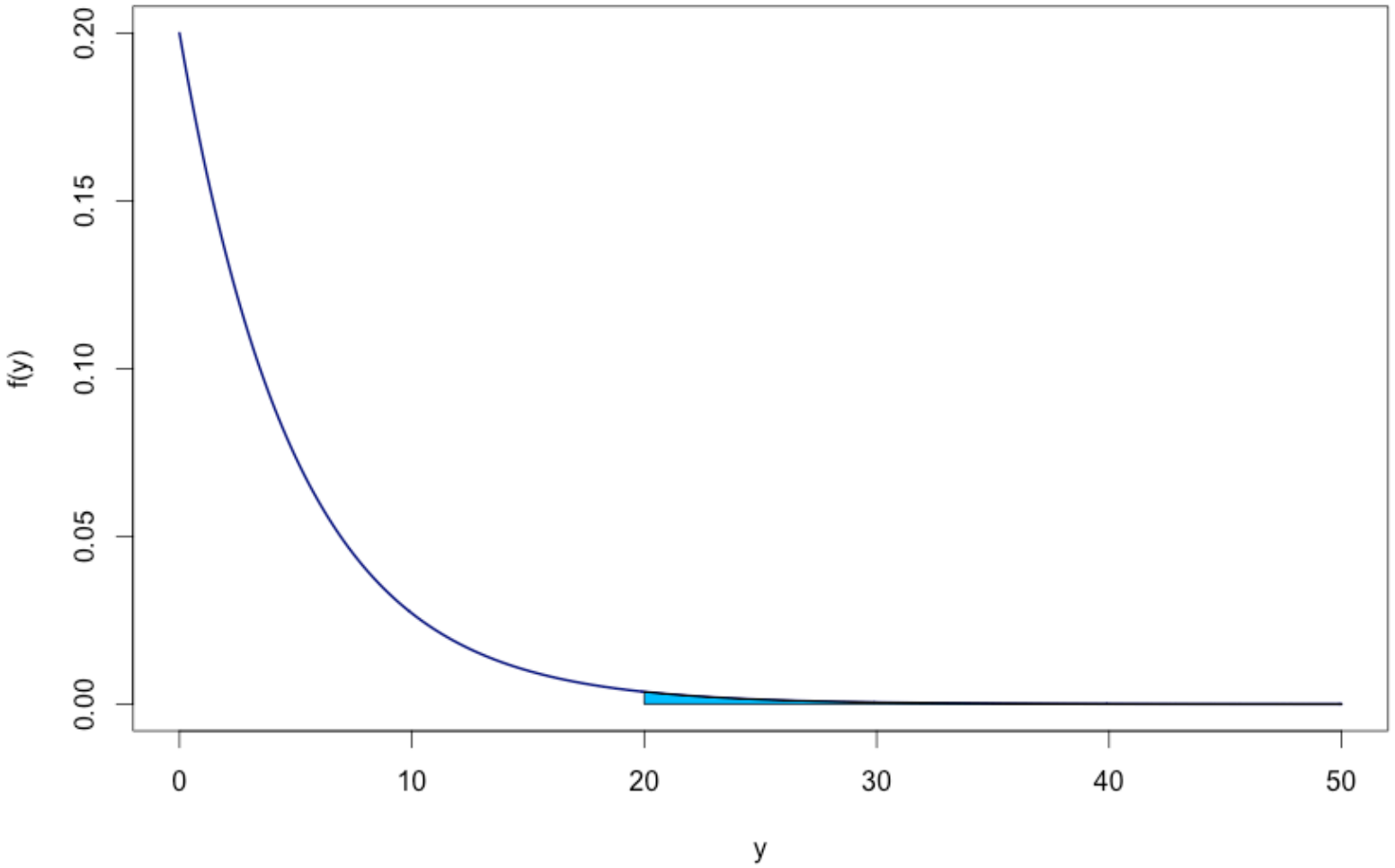
Random Variables & Distributions

$$P(Y \leq 10) = 0.8647$$



Random Variables & Distributions

$$P(Y > 20) = 0.0183$$



Random Variables & Distributions

- Probability distributions are efficiently summarized by a plots of the probability function, but it is also useful to characterized a probability function by a closed-form expression.
- Such expressions exist for several well-known probability distributions which are useful for describing a host of real-life random phenomenon.
- We discuss some such distributions, focusing on ones that are used routinely in the context of experimentation.

Random Variables & Distributions

The Binomial Distribution

$Y \sim BIN(n, \pi)$ counts the number of "success" in a sequence of n independent Bernoulli trials

$$f(y) = \binom{n}{y} \pi^y (1-\pi)^{n-y}$$

for $y = 0,1,2, \ldots, n$ and $\pi \in [0,1]$.

In R:

$P(Y = x)$ =`dbinom(x, size = 1, prob = p)`

DATA INSTITUTE
UNIVERSITY OF SAN FRANCISCO

Random Variables & Distributions

The Bernoulli Distribution

A $Y \sim BIN(1, \pi)$ random variable is said to have a Bernoulli Distribution. Thus, this is a special case of the binomial distribution when $n = 1$

$$f(y) = \pi^y (1 - \pi)^{1-y}$$

for $y = 0,1$ and $\pi \in [0,1]$.

In R:

$P(Y = x)$ =`dbinom(x, size = 1, prob = p)`

Random Variables & Distributions

The Normal Distribution

A $Y \sim N(\mu, \sigma^2)$ distribution has PDF given by

$$f(y) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-(y-\mu)^2}{2\sigma^2}}$$

for $-\infty < y < \infty$, $-\infty < \mu < \infty$ and $\sigma > 0$.

In R:

$P(Y \leq q) =$ `pnorm(q, mean = mu, sd = sigma)`

$P(Y > q) =$ `pnorm(q, mean = mu, sd = sigma, lower.tail = F)`

Random Variables & Distributions

The Standard Normal Distribution

This distribution arises as a special case of the normal distribution when $\mu = 0$ and $\sigma = 1$ and is typically denoted by $Z \sim N(0,1)$ with PDF

$$f(z) = \frac{1}{\sqrt{2\pi}} e^{\frac{-z^2}{2}}$$

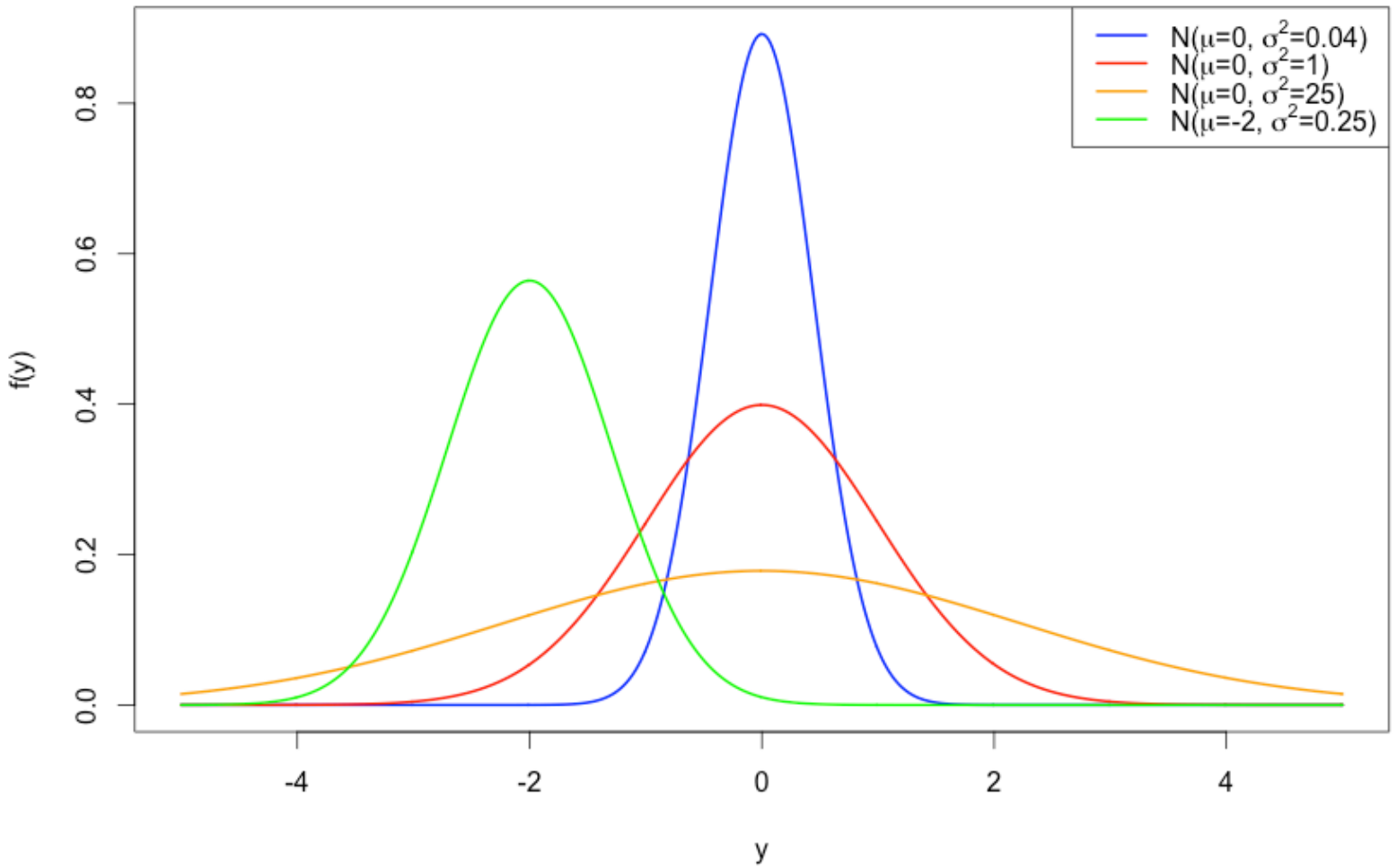
for $-\infty < z < \infty$.

In R:

$P(Z \leq q) =$ `pnorm(q)`

$P(Z > q) =$ `pnorm(q, lower.tail = F)`

Random Variables & Distributions



Random Variables & Distributions

The Student's *t*-Distribution

A $Y \sim t_{(\nu)}$ random variable has PDF given by

$$f(y) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu\pi}\,\Gamma\left(\frac{\nu}{2}\right)}\left(1+\frac{y^2}{\nu}\right)^{\frac{-(\nu+1)}{2}}$$

for $-\infty < y < \infty$ and $\nu > 0$.

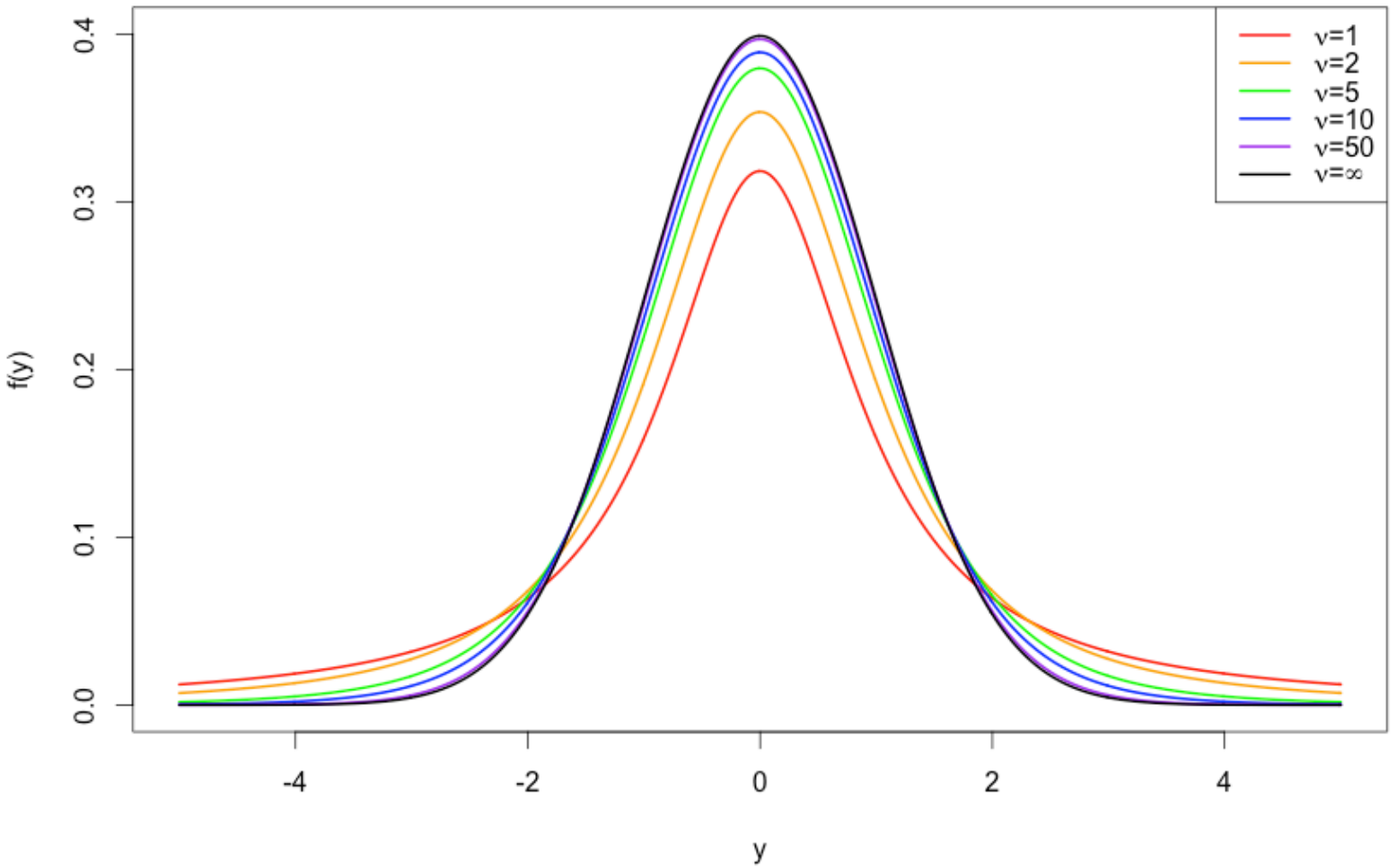
In R:

$P(Y \leq q) =$ `pt(q, df = v)`

$P(Y > q) =$ `pt(q, df = v, lower.tail = F)`

DATA INSTITUTE
UNIVERSITY OF SAN FRANCISCO

Random Variables & Distributions



Random Variables & Distributions

The Chi-Squared Distribution

A $Y \sim \chi^2_{(\nu)}$ random variable has PDF given by

$$f(y) = \frac{y^{\frac{k}{2}-1} e^{-y/2}}{2^{k/2}\Gamma\left(\frac{\nu}{2}\right)}$$

for $y \geq 0$ and a positive integer ν.

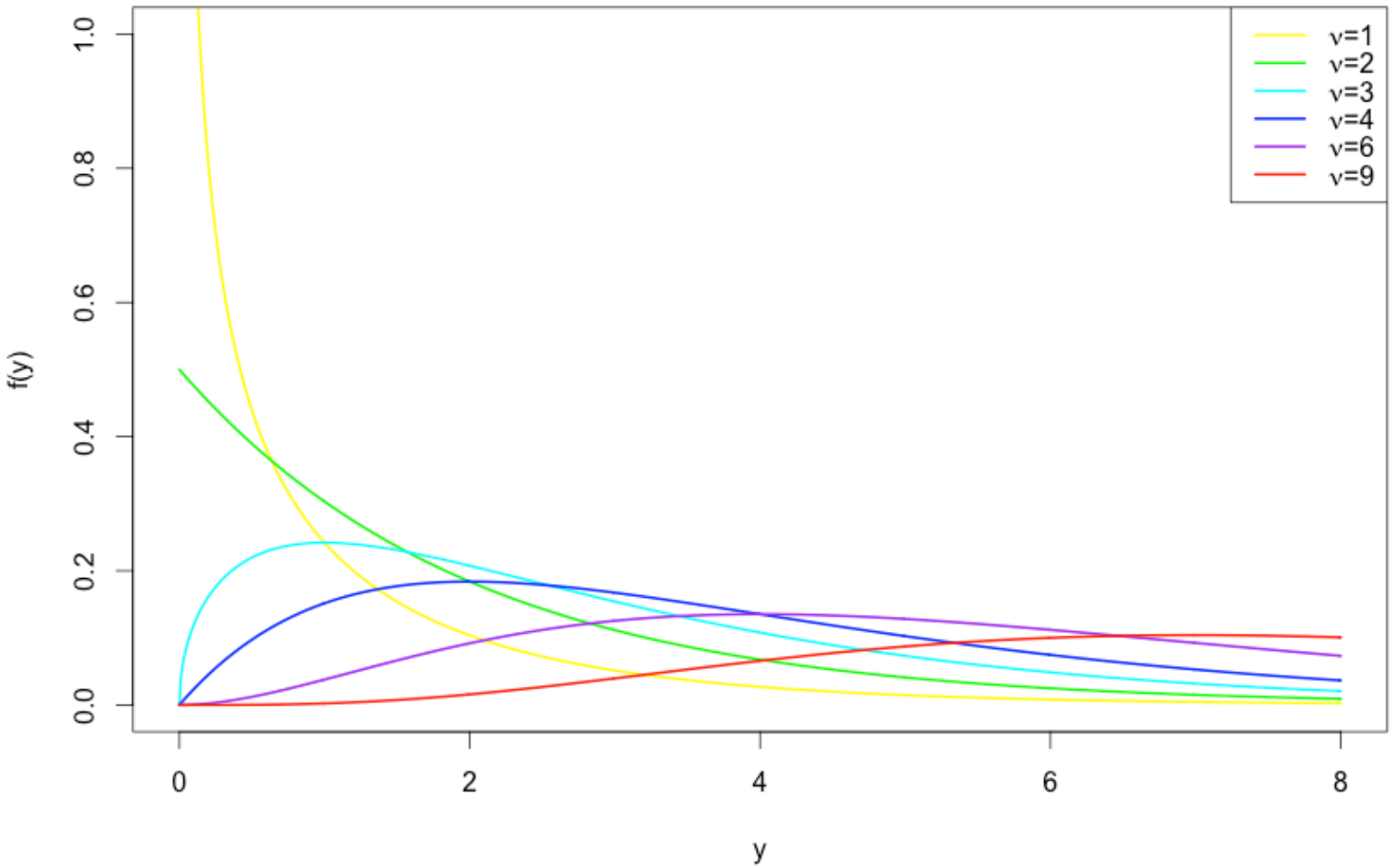
In R:

$P(Y \leq q) =$ `pchisq(q, df = v)`

$P(Y > q) =$ `pchisq(q, df = v, lower.tail = F)`

DATA INSTITUTE
UNIVERSITY OF SAN FRANCISCO

Random Variables & Distributions



Random Variables & Distributions

- The general shapes of these distributions are dictated by their PDFs, and by the parameters that govern them.
- However, the shape of ANY probability distribution, discrete or continuous, can be described by its moments.
- But to understand moments, we must first understand expected values.

Random Variables & Distributions

- The expected value of a random variable Y, denoted $\mathrm{E}[Y]$, is thought of as:
  - the "average" value of the random variable
  - the "most typical" outcome of the random process that governs Y
  - a measure of a probability distribution's "center"

- Mathematically, the expected value of Y is defined as:
  - $E[Y] = \sum_{all\ y} yf(y)$ if Y is discrete, and
  - $E[Y] = \int_{all\ y} yf(y)dy$ if Y is continuous, and

Random Variables & Distributions

- We then define the k^{th} moment of Y, or equivalently, the k^{th} moment of Y's distribution, as $\mathrm{E}[Y^k]$.
  - 1st moment: $\mathrm{E}[Y]$ quantifies center
  - 2nd moment: $\mathrm{E}[Y^2]$ quantifies spread
  - 3rd moment: $\mathrm{E}[Y^3]$ quantifies skewness
  - 4th moment: $\mathrm{E}[Y^4]$ quantifies kurtosis
- However, the first two moments are used most frequently in practice to describe a distribution's shape

Random Variables & Distributions

- While the 2nd moment $\mathrm{E}[Y^2]$ itself provides information about the dispersion of a distribution, it is most commonly used in the calculation of the variance of Y:

$$Var[Y] = E[(Y - E[Y])^2]$$
$$= E[Y^2] - E[Y]^2$$

- The variance is interpreted as the expected squared deviation from the mean

- Note that the dispersion of a distribution is also commonly communicated in terms of the standard deviation of Y:

$$SD[Y] = \sqrt{Var[Y]}$$

Random Variables & Distributions

Distribution	$E[Y]$	$Var[Y]$
$Y \sim BIN(n, \pi)$	$n\pi$	$n\pi(1-\pi)$
$Y \sim N(\mu, \sigma^2)$	μ	σ^2
$Y \sim t_{(\nu)}$	0	$\nu/(\nu-2)$
$Y \sim \chi^2_{(\nu)}$	ν	2ν

DATA INSTITUTE
UNIVERSITY OF SAN FRANCISCO

Statistical Inference

In real life we do not observe probability models, we observe data that have been sampled from some population.

Statistical Inference occurs when one uses sample data to draw conclusions about the population from which the sample was drawn.

If a population can reasonably be modeled by some well-studied probability distribution, then drawing conclusions about it is straightforward.

Statistical Inference

- As we saw previously, much of distribution's information is contained in its shape, and the shape of a given distribution relies entirely on one or more parameters.
- More generally, we think of all statistical models as being governed by one or more unknown parameters.
- Because parameter values are unknown in practice interest lies in
    i. Estimating these parameters in light of the observed data
    ii. Testing hypotheses about the parameters

DATA INSTITUTE
UNIVERSITY OF SAN FRANCISCO

Estimation

When a data scientist says that they are fitting a model to some data, what they really mean is:

- They've assumed a certain model or probability distribution is appropriate for describing some characteristic or relationship in a population.
- They have collected data (i.e., a sample from the population) with which they intend to study this characteristic or relationship.
- They intend to use the observed data to estimate the unknown parameters associated with the model or distribution.

Estimation

The goal of point estimation is to use observed data to obtain reasonable values of a model's unknown parameters that are consistent with the data that were actually observed.

We typically use Greek letters, like θ, to denote parameters and we use $\hat{\theta}$ to denote an estimate of the parameter calculated using sample data.

The notation $\hat{\theta}$ is read as “theta-hat”

Estimation

In general, a variety of estimation methods may be used to obtain parameter estimates:

- The Method of Moments
- Maximum Likelihood Estimation
- Least Squares Estimation.
- Etc….

All estimation procedures have advantages and disadvantages, and so it is important to choose the one that is appropriate for your data and your problem.

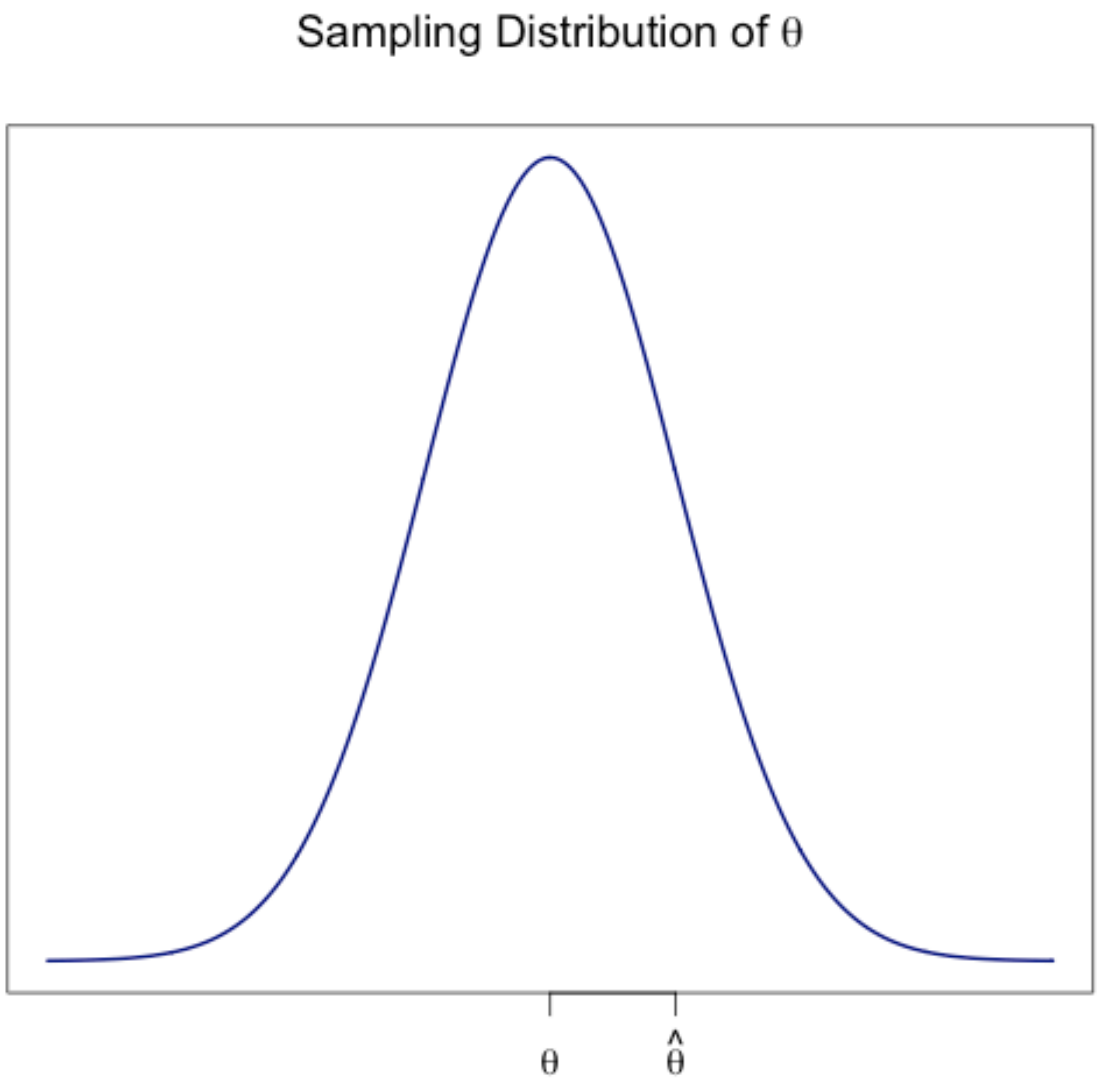
Estimation

However, there is no guarantee that $\hat{\theta}$ is anywhere close to the true value of θ.

A point estimate of θ is just a single value, based on one sample of data.

If we were to draw a second sample and repeat the exact same estimation procedure we would very likely obtain a slightly different value of $\hat{\theta}$ than before, simply due to sampling variation.

Estimation



Estimation

The point estimate $\hat{\theta}$ communicates no information about the sampling distribution on its own.

For this reason we typically accompany point estimates with interval estimates, also know as conference intervals.

A confidence interval accounts for sampling variation and provides an interval within which we can be confident the true value of θ lies.

Thus a point estimate provides the **best estimate** of θ based on the observed data and a confidence provides a **range of plausible values** for θ accounting for sampling variation.

DATA INSTITUTE
UNIVERSITY OF SAN FRANCISCO

Estimation

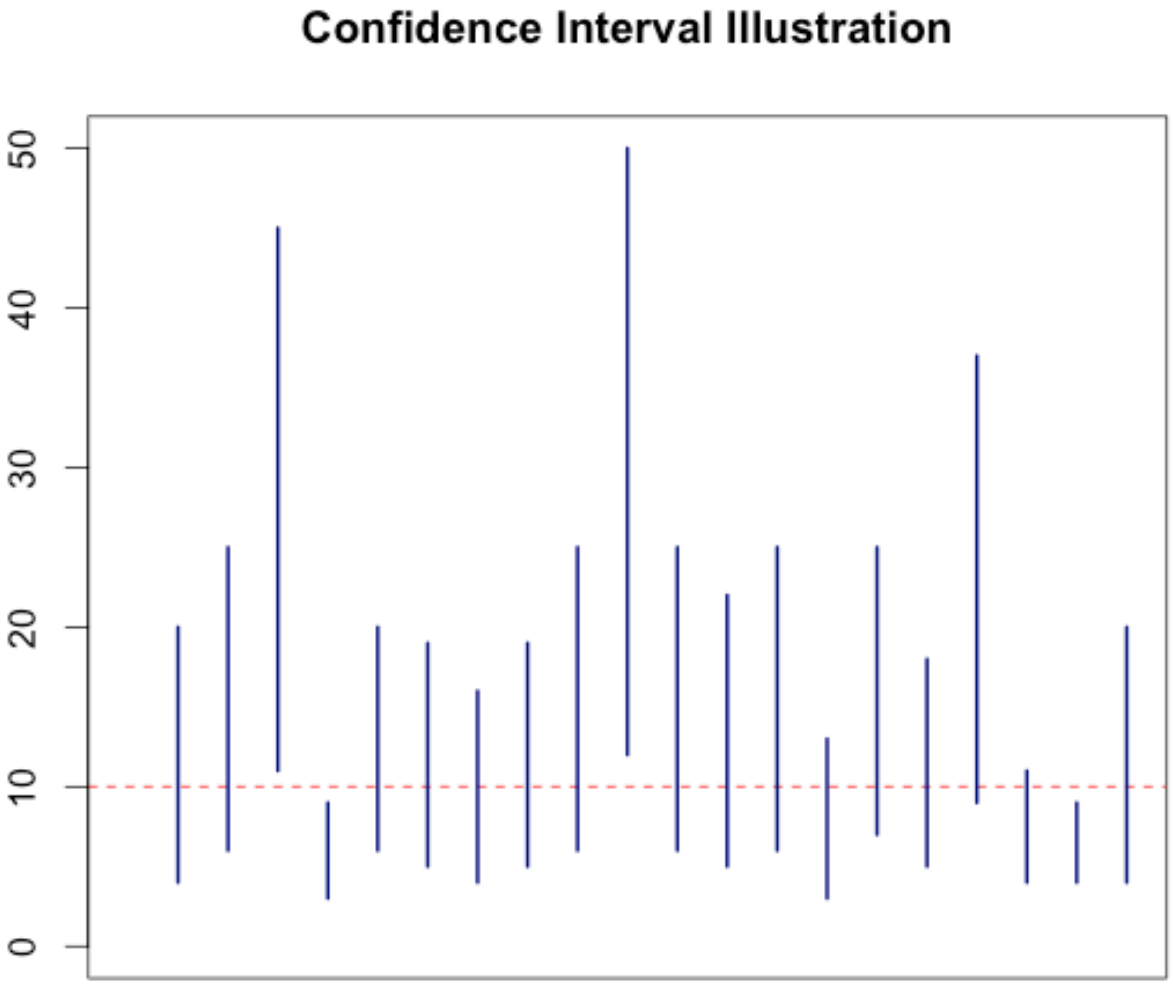
The most common confidence interval for a parameter θ is a 95% confidence interval.

We interpret such intervals by saying that **we are 95% confident** that the true value of θ is contained in this interval.

But what does it mean to be "95% confident"?

What this means is that if we were to make infinitely many such intervals, 95% of them would contain θ.

Estimation



Hypothesis Testing

With estimation we knew nothing about the parameter of interest, so we estimated it from observed data.

However, we may have a belief as to what θ is, and if we'd like to formally evaluate this belief, we should perform a hypothesis test.

Formally, a hypothesis is a statement about a parameter that we'd either like to prove or disprove, using the collected data.

A hypothesis is tested by comparing ones own data with a hypothesized statistical distribution.

DATA INSTITUTE
UNIVERSITY OF SAN FRANCISCO

Hypothesis Testing

We typically state such a the hypothesis as

$$H_0: \theta = \theta_0 \text{ versus } H_A: \theta \neq \theta_0$$

where

- H_0 is the null hypothesis, and it is the statement we believe to be true, and that we want to test using observed data.
- H_A is the alternative hypothesis, and it is typically that which we would like to prove.

Exactly one of these hypotheses is true, and we use observed data to try and empirically uncover the truth

DATA INSTITUTE
UNIVERSITY OF SAN FRANCISCO

Hypothesis Testing

The previous hypothesis statement corresponded to a two-sided hypothesis.

The following statements correspond to one-sided hypotheses

$$H_0: \theta \leq \theta_0 \text{ versus } H_A: \theta > \theta_0$$

or

$$H_0: \theta \geq \theta_0 \text{ versus } H_A: \theta < \theta_0$$

Which hypothesis statement is appropriate depends on the context of the problem and the question that the hypothesis test is designed to answer

DATA INSTITUTE
UNIVERSITY OF SAN FRANCISCO

Hypothesis Testing

Formally, we use observed data to decide whether to reject or not reject H_0.

In order to draw such a conclusion, we define a test statistic T which is a random variable that satisfies three properties

- it must be a function of the observed data
- it must be a function of the parameter θ
- its distribution must not depend on θ

Then, assuming the null hypothesis is true, the test statistic T follows a particular distribution which we call the null distribution

Hypothesis Testing

Next we calculate t, the observed value of the test statistic, by substituting the observed data and θ_0 into the expression for T.

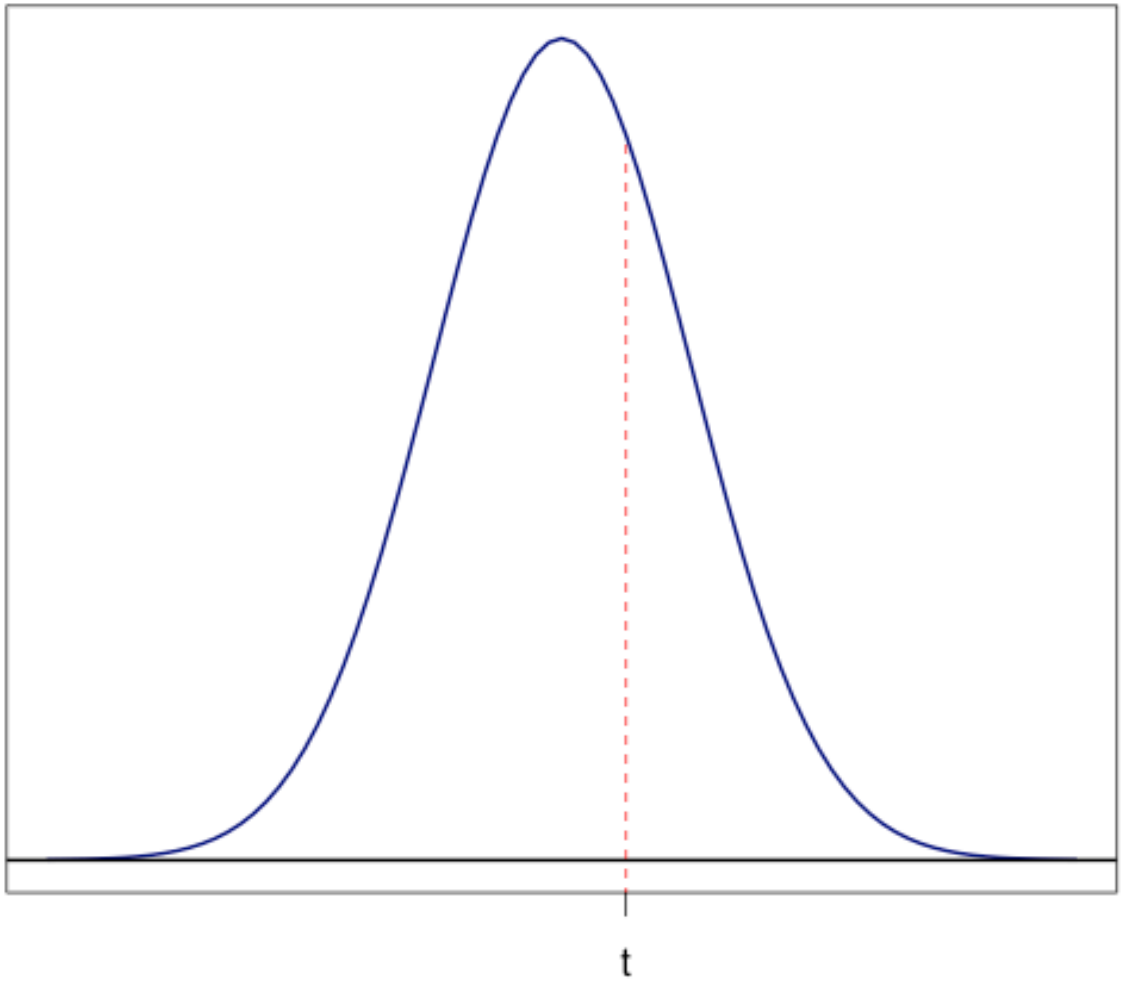
Note that expressions for T typically incorporate terms of the form:

$$\hat{\theta} - \theta_0 \qquad \text{or} \qquad \hat{\theta}/\theta_0$$

Then we evaluate the extremity of t relative to the null distribution:

- If t seems very extreme, this provides evidence against H_0
- If t seems reasonable, this provides evidence in favor of H_0

Hypothesis Testing



Hypothesis Testing

We formalize the extremity of t using the p-value of the test.

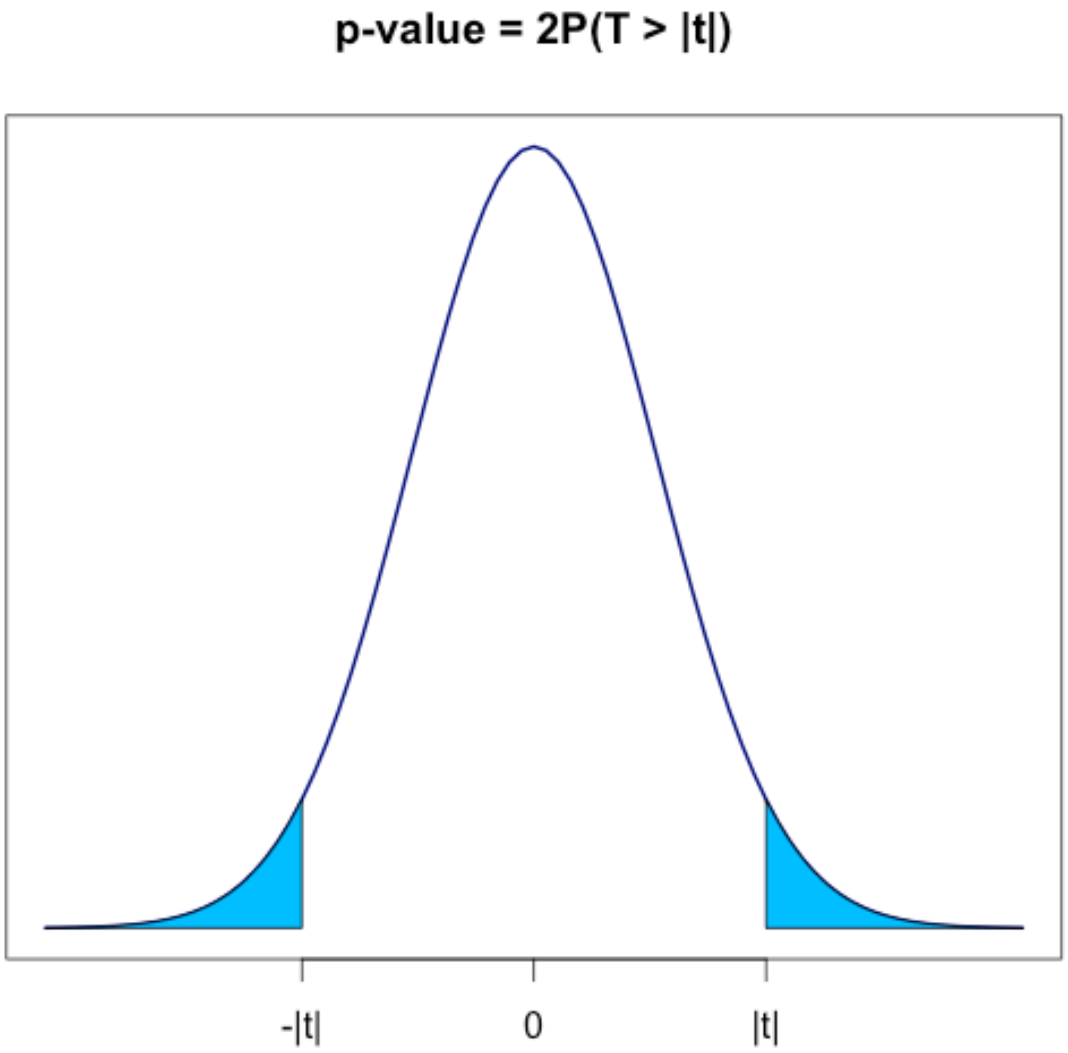
The p-value is defined to be:

The probability of observing a value of the test statistic at least as extreme as the value we observed, if the null hypothesis is true.

Whether large values of t, small values of t, or both, are to be considered extreme depends on whether H_A is one- or two-sided

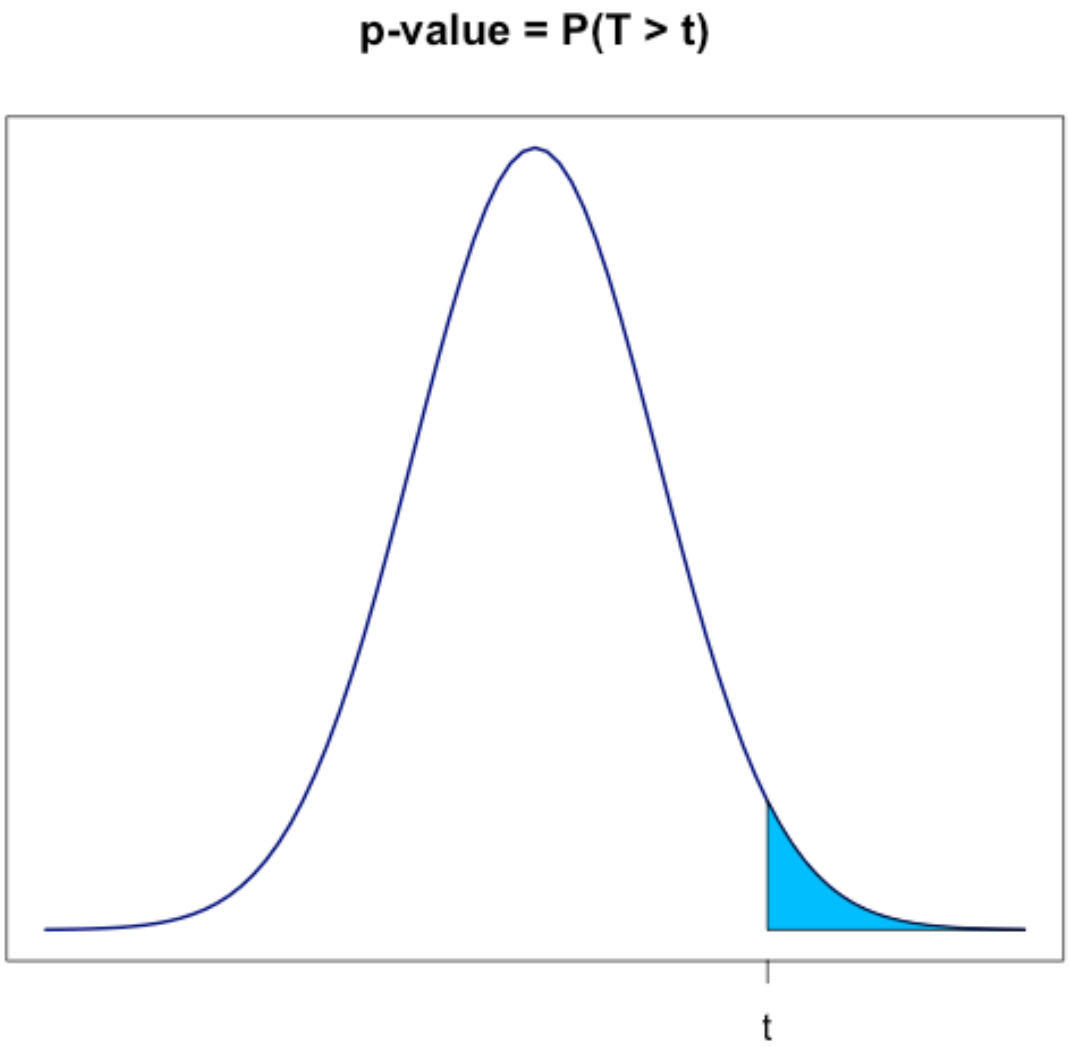
Hypothesis Testing

$H_0: \theta = \theta_0$ versus $H_A: \theta \neq \theta_0$



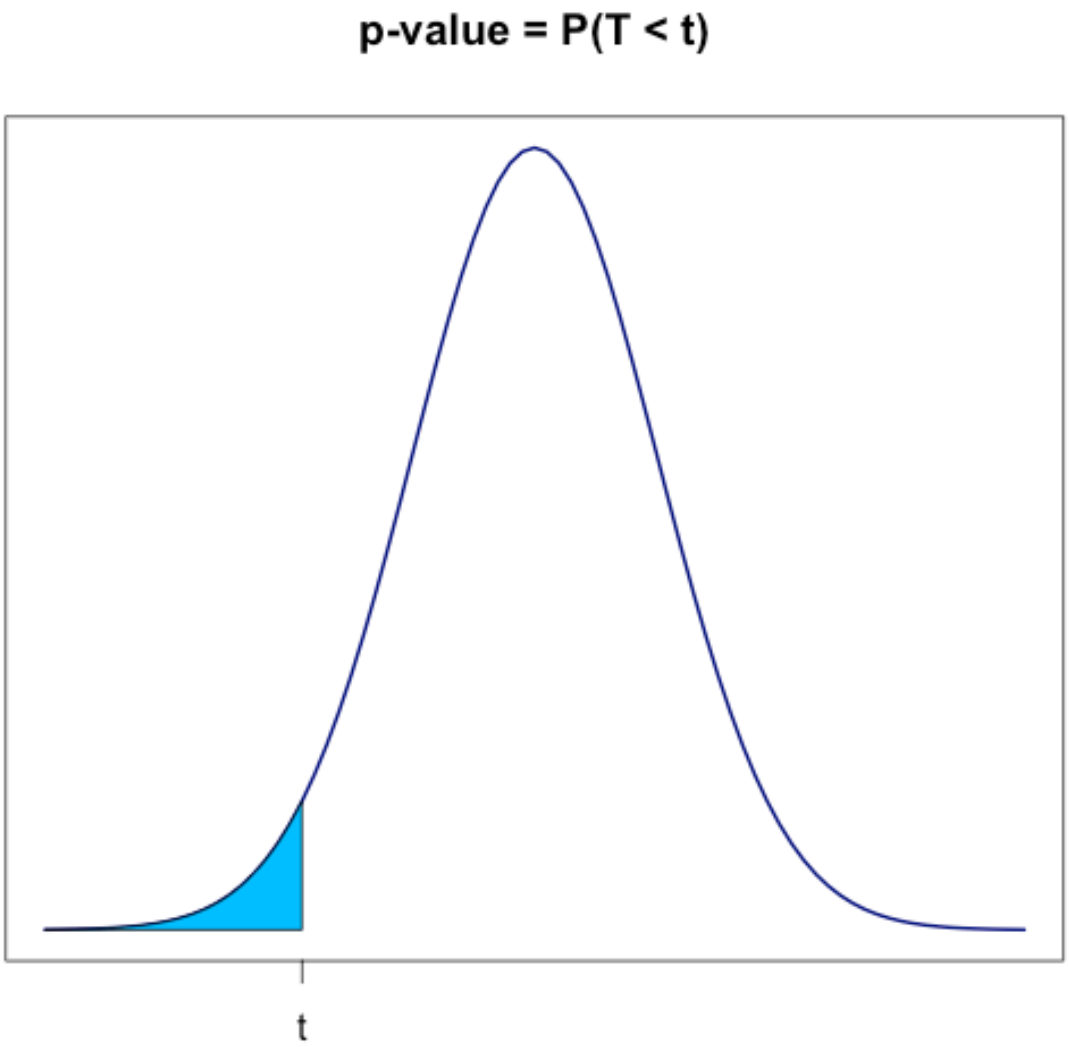
Hypothesis Testing

$H_0: \theta \leq \theta_0$ versus $H_A: \theta > \theta_0$



Hypothesis Testing

$H_0: \theta \geq \theta_0$ versus $H_A: \theta < \theta_0$



Hypothesis Testing

How "extreme" t must be, and hence how small the p-value must be, to reject H_0 is determined by the significance level of the test, which we denote by α.

In particular, if

- p-value $\leq \alpha$ we reject H_0 in favor of H_A
- p-value $> \alpha$ we do not reject H_0

Note that $\alpha = 0.01$ and 0.05 are common choices.

Hypothesis Testing

In drawing such a conclusion, it is possible that we make one of two types of errors:

- Type I Error: based on the observed data we reject H_0 when it is in fact true
- Type II Error: based on the observed data we fail to reject H_0 when it is in fact false.

It is important to note that there is typically an imbalance in the consequences associated with these two types of error.

Hypothesis Testing

Courtroom Analogy

H_0: *defendant is innocent* vs. H_A: *defendant is guilty*

- Type I Error: the defendant is truly innocent, but the evidence leads the jury to find the defendant guilty
- Type II Error: the defendant is truly guilty, but the evidence leads the jury to find the defendant innocent

In any hypothesis testing setting, both types of errors lead to negative outcomes, but these negative outcomes may be prioritized differently

DATA INSTITUTE
UNIVERSITY OF SAN FRANCISCO

Hypothesis Testing

Clearly we would like to reduce the likelihood of either type of error happening.

We define

$$\alpha = P(\text{Type I Error}) \text{ and } \beta = P(\text{Type II Error})$$

which reflect the chances that a Type I or Type II error will occur.

We call α the significance level of the test and $1 - \beta$ the power of the test.

Thus a test with a **small significance level** and **large power** is desirable.

DATA INSTITUTE
UNIVERSITY OF SAN FRANCISCO

Hypothesis Testing

In practice we choose α and β based on how often we are comfortable allowing Type I and Type II errors to occur.

Suppose we can tolerate a Type I error 1% of the time and a Type II error 5% of the time.

In this case we would choose $\alpha = 0.01$ and $\beta = 0.05$ and the significance level of the test would be 5% and the power of the test would be 95%

Common choices for significance level and power are respectively 5% and 80%, corresponding to $\alpha = 0.05$ and $\beta = 0.2$

Hypothesis Testing

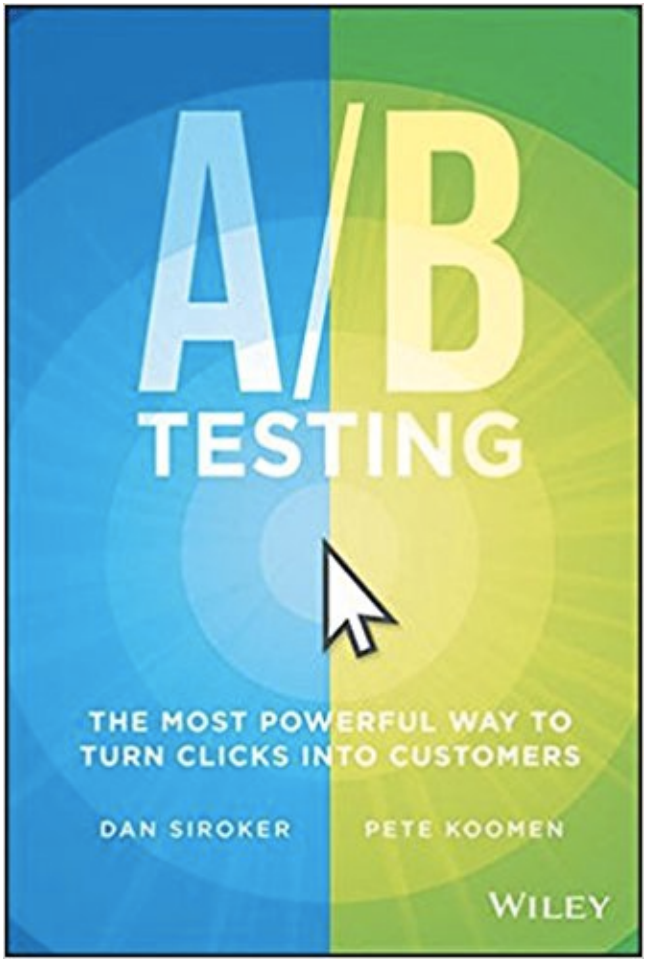
Thus the significance level α (i.e., the probability of making a Type I error), determines how small a p-value must be in order to reject a null hypothesis.

Next week we will see that for a fixed value of α the desired power determines the required sample size.

The choices for α and β should be made prior to testing the hypothesis and in fact prior to collecting any data.

Take Home Task

Read this book!



See you next week!

DATA INSTITUTE
UNIVERSITY OF SAN FRANCISCO