Probability Distributions

BRSM

The role of assumptions in statistics

Before the match, Fischer had won 3 games, Taimanov had won 2 games, and 1 game was drawn.

We bet on the winner of the next game, after each round.

The limits of logic in everyday life.

What is statistical inference?

- Polling company
- Randomly call 1000 people
- 35% said they'd vote for XYZ party
- The result comes out. The number actually is 26%
- The question is: how surprised (or not) should we be by this result?
- To do this, we need tools for statistical inference
- Each tool makes some assumptions about the data
- We need to understand probabilities and probability distributions first

What is the difference between probability and statistics?

- What is the probability that in two successive coin tosses, you get both tails?
- You have the model of the world here (e.g. it is a fair coin, $P(H) = 0.5$), but no data and are asked to come up with the probability of a hypothetical event
- Going back to Fischer-Taimanov, after 3 rounds and 3 wins to Fischer, we are to make an inference about what model is correct, given the 3 win data. Is P(Fischer) really 0.5 or is it something else? This is the realm of inferential statistics.

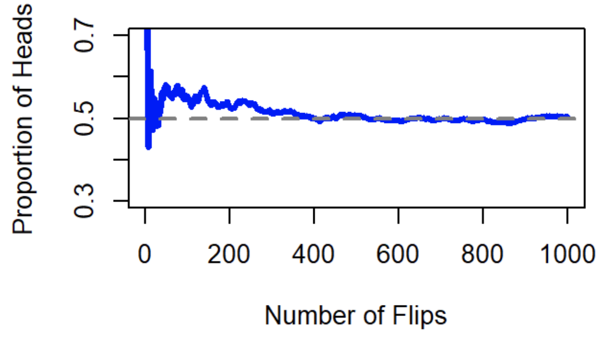
What is a probability?

- Means slightly different things if you are a frequentist statistician vs if you are a Bayesian
- Carlsen has a 70% chance of winning a game against Nepomniachtchi: what does this mean to you?
- If they play a 10 game match, Carlsen is expected to win 7?
- If I bet Rs 100 on Nepomniachtchi, I should get a reward of Rs 233 (700/3) if Nepo wins against your bet of Rs 233 on Carlsen (and if Carlsen wins, you get Rs 100).
- 70% reflects my subjective belief of how much stronger Carlsen is compared to Nepo.

Frequentist probability

FLIP A COIN MANY TIMES AND COUNT THE PROPORTION OF HEADS

- As N --> infinity, the probability converges to the true probability
- Frequentist statistics rely on assumptions about how you sample the data (just like a coin toss), and cares about long-run proportions of a certain result (e.g. heads) in such hypothetical future samples.



Frequentist statistics

- Pros: objective because anyone following the same "sampling plan" will observe a similar proportion over the long run.
- Cons: The equivalent of flipping a coin infinite times to understand a probability can be counterintuitive in practice: "There is 80% chance of rain today." We can intuitively somehow understand what this means.
- The interpretation in frequentist terms: "There is a class of day for which if we observe across N--> infinite days, it rained on 80% of those days".
- This type of conundrum is exactly what you will see drives debates in statistical methods between frequentists and Bayesians.

Bayesian probability

Subjective

Minority view amongst statistical practitioners

Degree of subjective belief assigned to an event

Bayesian probability

- Pros:
  - You can assign probabilities to non-repeatable events
  - You can legitimately interpret the probability as degree of belief (similar probabilities in the frequentist world will have more convoluted interpretations leading to the sorts of pitfalls we discussed/will discuss about p-values, confidence intervals, etc).
- Cons:
  - Not objective
  - Depends on priors (background knowledge), which can be subjective

Independent Events

- Two events A and B are independent if
- $P(AB) = P(A).P(B)$
- $P(A|B) = P(AB)/(P(B)) = P(A)$

Variables and their distributions

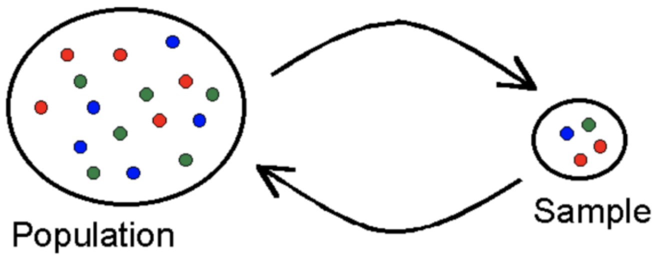
- You will often hear things like "variable x is i.i.d"
- Independently and identically distributed
- Say Yi are dice throws for i=1:n
- The outcome of each different set of (n throws) is a random variable itself
- The outcome of each throw has the same distribution (uniform over 6 possibilities): Y1, Y2, …,Yn are identically distributed
- Y1 is independent of Y2 and so on.
- Therefore, iid.

A function applied on the sample

- Yi is iid
- Now, if we apply a function on the sample, such as a sum or an average, this is also a random variable
- We can also talk about distributions of such variables!
- This is an important concept in statistics: **sampling distribution of some statistic**

Sample vs population

- Sample (data sample) : e.g. one particular "sample" of N throws or one particular sample of 1000 people in an exit poll in Punjab
- Population: e.g. The universal set of all possible N throw outcomes or all voters in Punjab



Distribution of what? Be clear

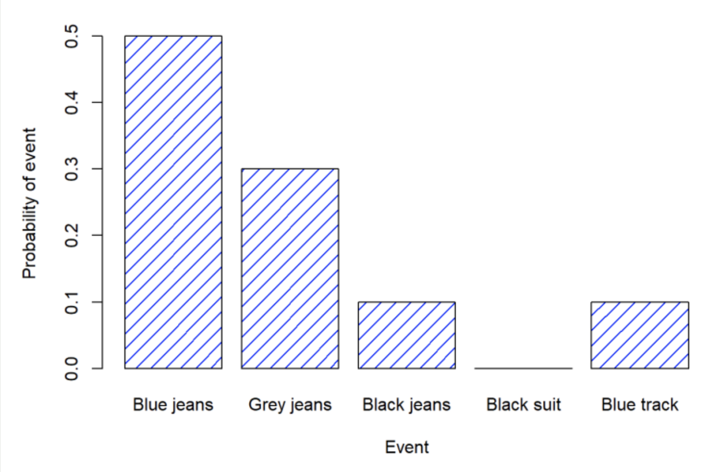
SAMPLING DISTRIBUTION OF A STATISTIC: THE DISTRIBUTION OF A STATISTIC (OR A FUNCTION) APPLIED ON THE SAMPLES

POPULATION: WHAT IS THE DISTRIBUTION OF VOTING PREFERENCES TAKEN FROM THE ENTIRE POPULATION OF PUNJAB?

NEED TO BE CLEAR ABOUT THE DISTINCTIONS

Probability distribution

Which.pants	Blue.jeans	Grey.jeans	Black.jeans	Black.suit	Blue.tracksuit
Label	X_1	X_2	X_3	X_4	X_5
Probability	$P(X_1) = .5$	$P(X_2) = .3$	$P(X_3) = .1$	$P(X_4) = 0$	$P(X_5) = .1$



Probability density function (PDF)

$$P(a \leq X \leq b) = \int_a^b f(x)\,dx.$$

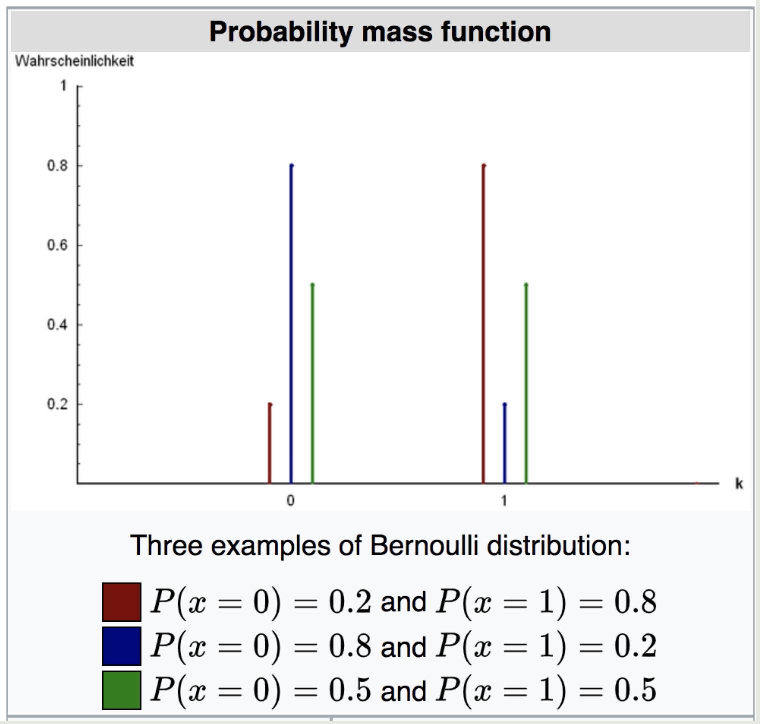
Defined for continuous random variables
The probability that x = an exact value = 0 for continuous variables because a = b in this integral

Cumulative Distribution Function (CDF)

$$F(x) = P(X \leq x) = \int_{-\infty}^{x} f(t)\, dt$$

Discrete variables: Bernoulli Distribution

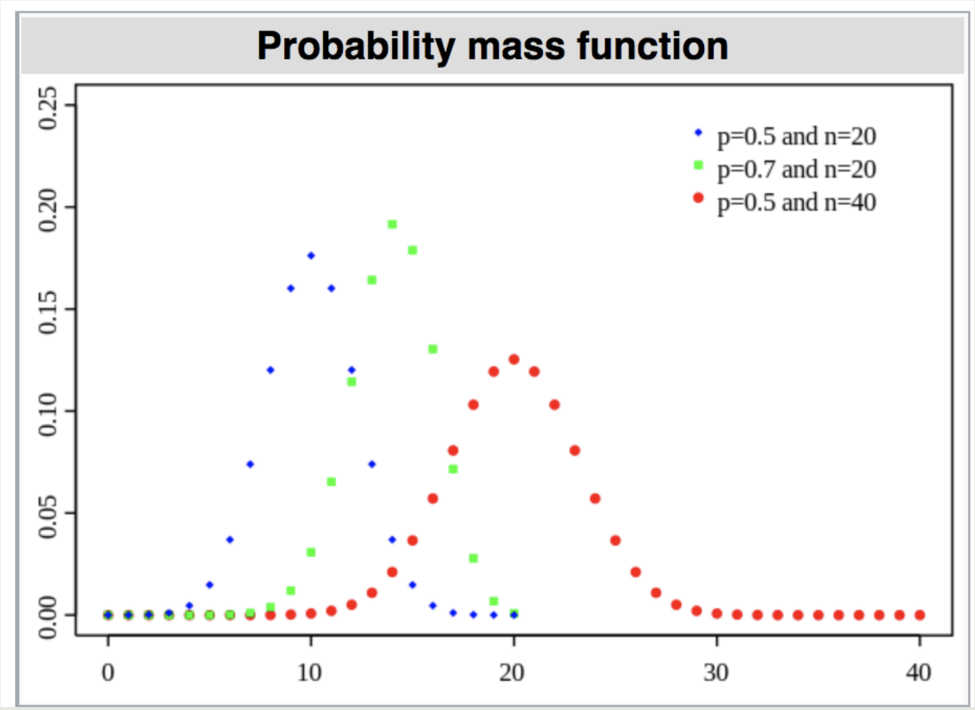
- The Bernoulli distribution is the discrete probability distribution of a random variable which takes a binary, boolean output: 1 with probability p, and 0 with probability (1-p).



Wikipedia

Binomial distribution

- If there is a series of n i.i.d Bernoulli trials (all trials have a success probability of p), then the sum of outcomes is distributed as Binom(n,p)



Wikipedia

Notation

$$X \sim \text{Binomial}(\theta, N)$$

Working with distributions in R

Table 9.3: The naming system for R probability distribution functions. Every probability distribution implemented in R is actually associated with four separate functions, and there is a pretty standardised way for naming these functions.

What.it.does	Prefix	Normal.distribution	Binomial.distribution
probability (density) of	d	dnorm()	dbinom()
cumulative probability of	p	dnorm() pnorm()	pbinom()
generate random number from	r	rnorm()	rbinom()
q qnorm() qbinom()	q	qnorm()	qbinom(

What is the probability of observing 6 heads in 10 coin tosses given an unfair coin?

- P = 0.7
- **dbinom**(x = 6, size = 10, prob = 0.7)
- 0.2001209

R distributions

The d form we've already seen: you specify a particular outcome x, and the output is the probability of obtaining exactly that outcome. (the "d" is short for ***density***, but ignore that for now).

The p form calculates the ***cumulative probability***. You specify a particular value q, and it tells you the probability of obtaining an outcome *smaller than or equal to* q.

The q form calculates the ***quantiles*** of the distribution. You specify a probability value p, and gives you the corresponding percentile. That is, the value of the variable for which there's a probability p of obtaining an outcome lower than that value.

The r form is a ***random number generator***: specifically, it generates n random outcomes from the distribution

10 coin tosses

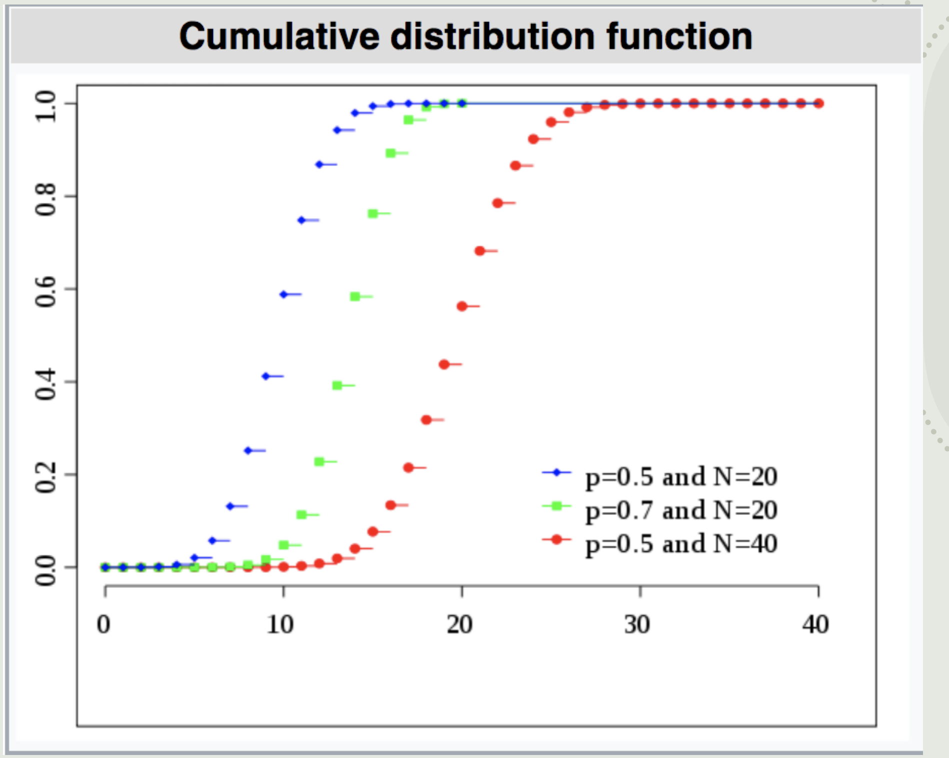
- Probability that I get <= 4 heads?
- P(1) + P(2) + P(3) + P(4) = **dbinom**(x = 1, size = 10, prob = 0.7) + **dbinom**(x = 2, size = 10, prob = 0.7) + **dbinom**(x = 3, size = 10, prob = 0.7) + **dbinom**(x = 4, size = 10, prob = 0.7)
- 0.04734308

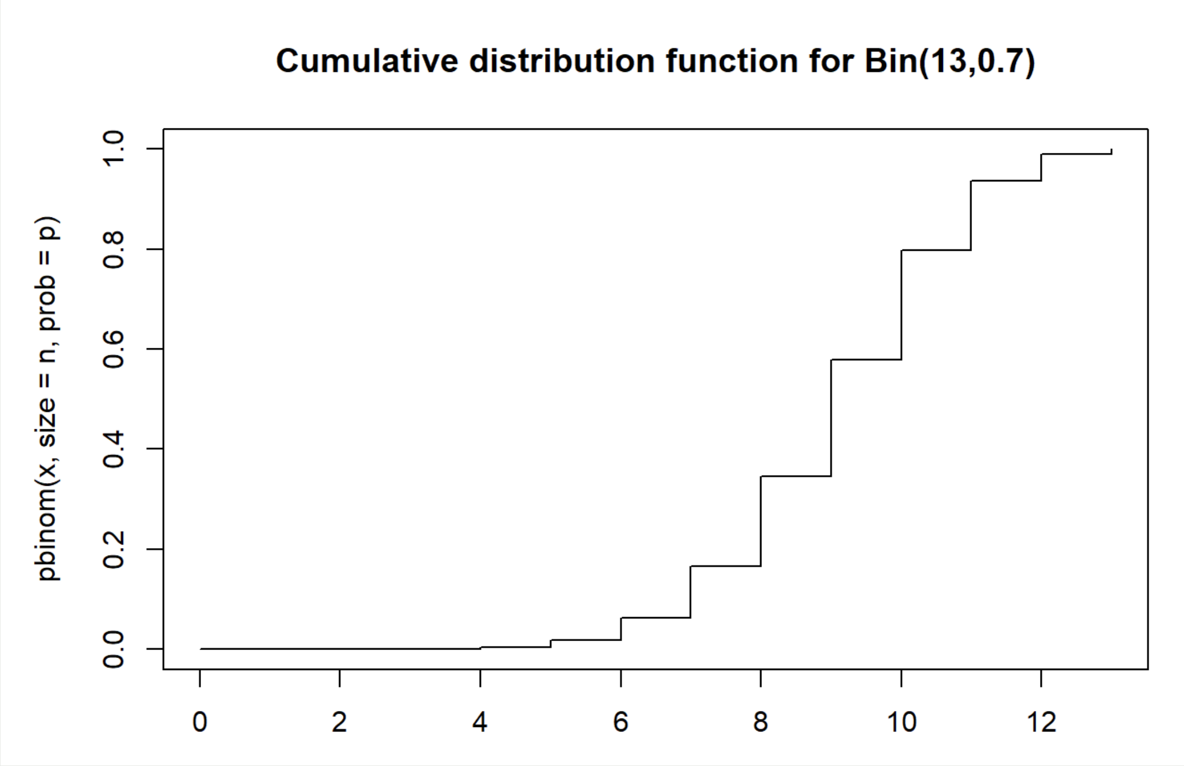
- Easier way: **pbinom**(q= 4, size = 10, prob = 0.7)
- 0.04734899 (4 is the 4.7 th percentile of the Binomial data or 4.7% of the values fall under 4)

- qbinom(p = 0.04, size = 10, prob = 0.7)
- 4 (the 4 th percentile of the data is 4)
- Wait, how can the 4th percentile also be 4??
- The Binomial distribution here doesn't really have a 4th percentile.

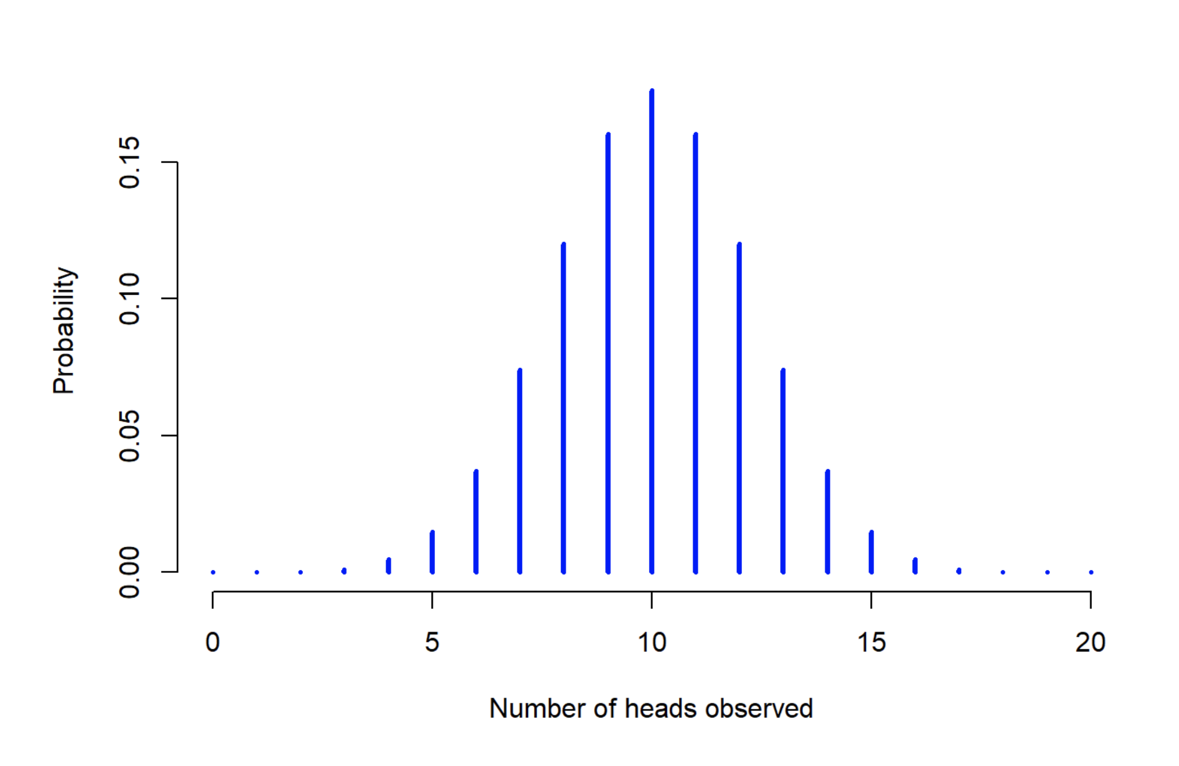
Warning: discrete variables and cumulative distribution functions

- Supported only on countable numbers
- So only some percentiles on the Y axis ->
- If you provide it any other percentile, the R function will round upwards.
- Not a problem for continuous distributions

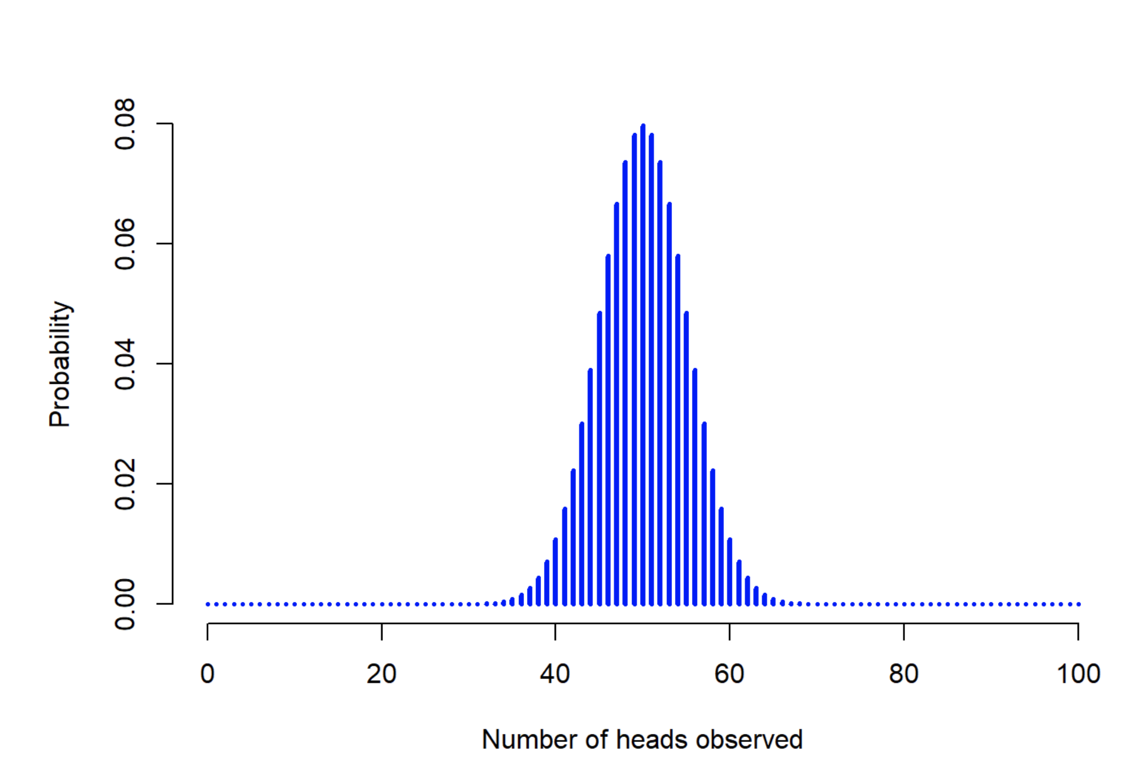



Cumulative distribution function for Bin(13,0.7)
pbinom(x, size = n, prob = p)
0.0
0.2
0.4
0.6
0.8
1.0
0
2
4
6
8
10
12

Flip a fair coin 20 times



Flip a fair coin 100 times

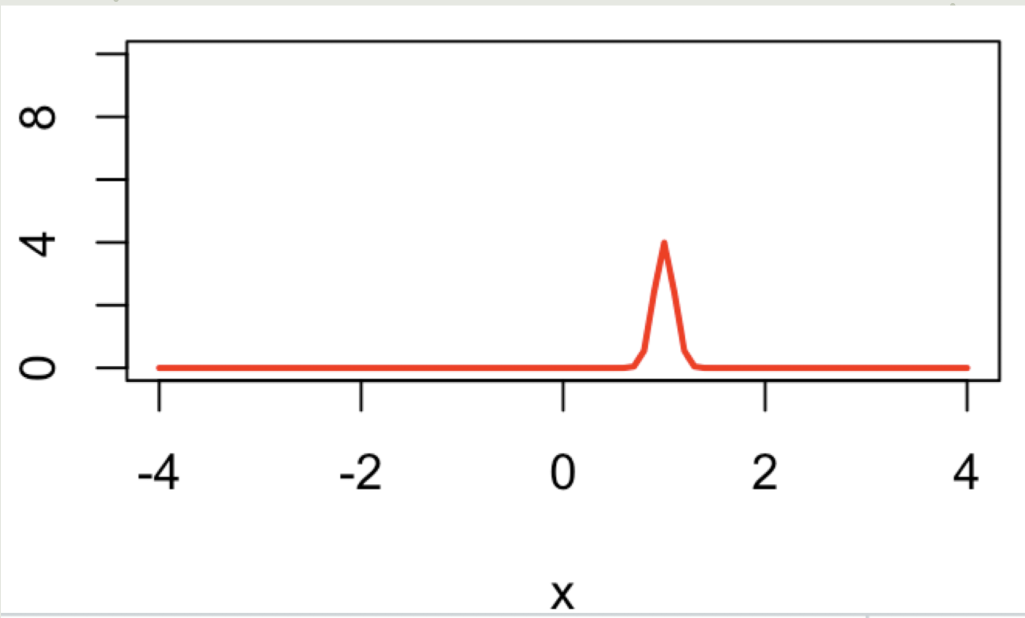


Normal Distribution

$$X \sim \mathrm{Normal}(\mu, \sigma)$$

Normal
$p(X\|\mu,\sigma) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{(X-\mu)^2}{2\sigma^2}\right)$

```
plot(x, dnorm(x, mean = 1, sd = 0.1), type = "l",
    ylim = c(0, 10), ylab = "", lwd = 2, col = "red")
```

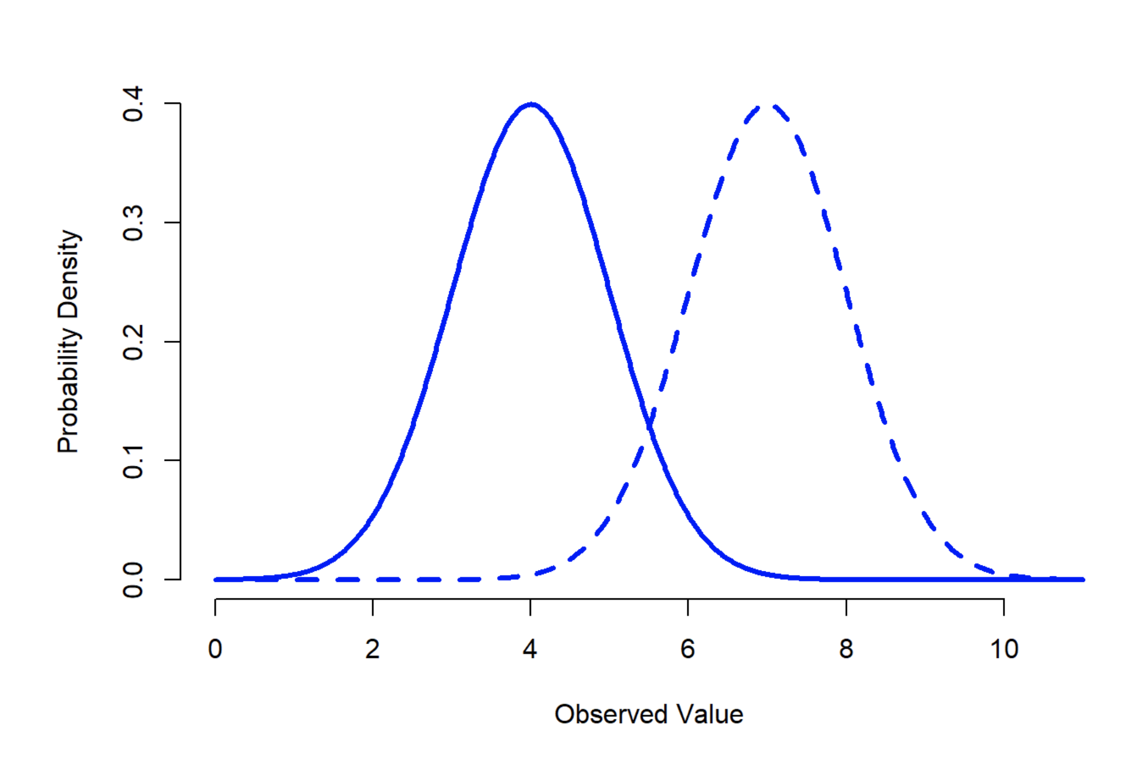


Normal PDF

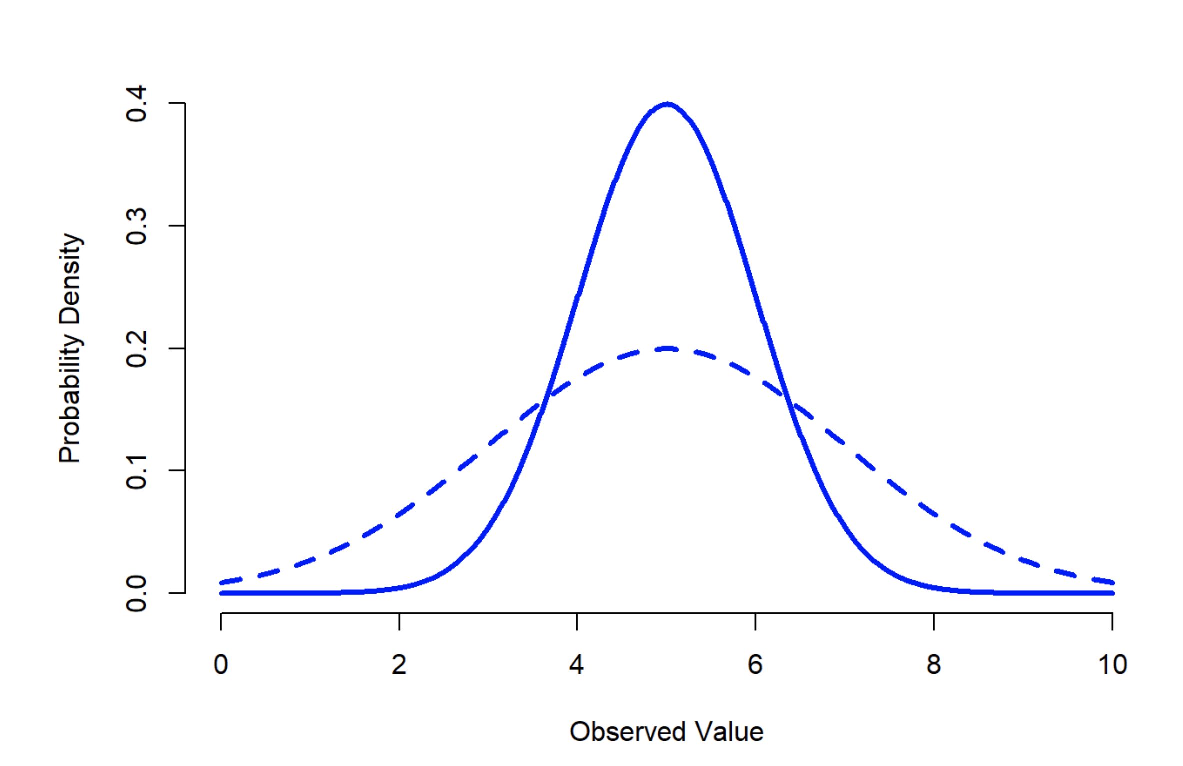
Q: What is the probability that x = 1?

```
> dnorm( x = 1, mean = 1, sd = 0.1 )
[1] 3.989423
```

Different means, same standard deviation ("width")



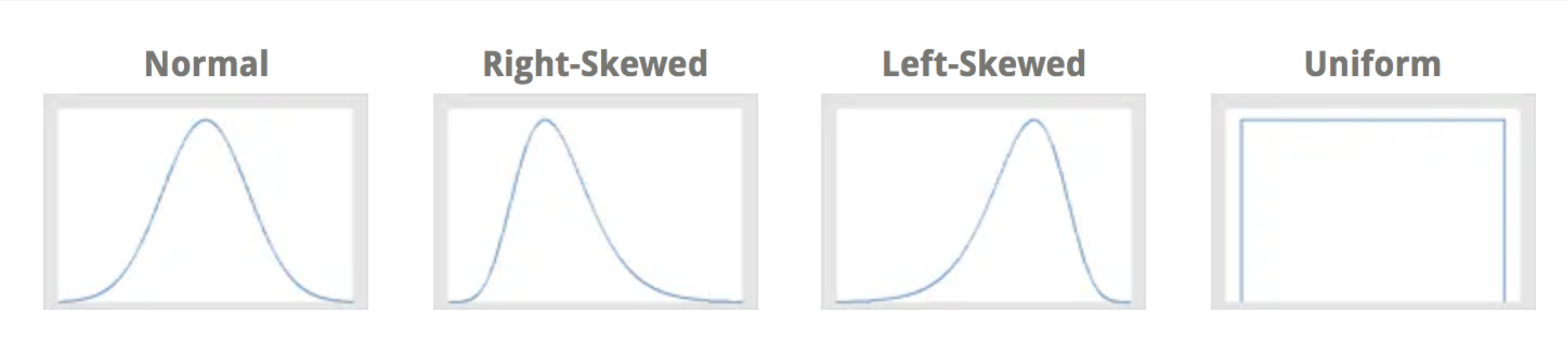
Same mean, different widths



Central Limit Theorem

- The central limit theorem states that, given a sufficiently large sample size, the sampling distribution of the mean for a variable will approximate a normal distribution regardless of that variable's distribution in the population.

Applies to almost all probability distributions of the population



The above is the distribution of the variable in the population!
Now you draw a random sample of size n from this.

The only requirement: the population distribution must have finite variance

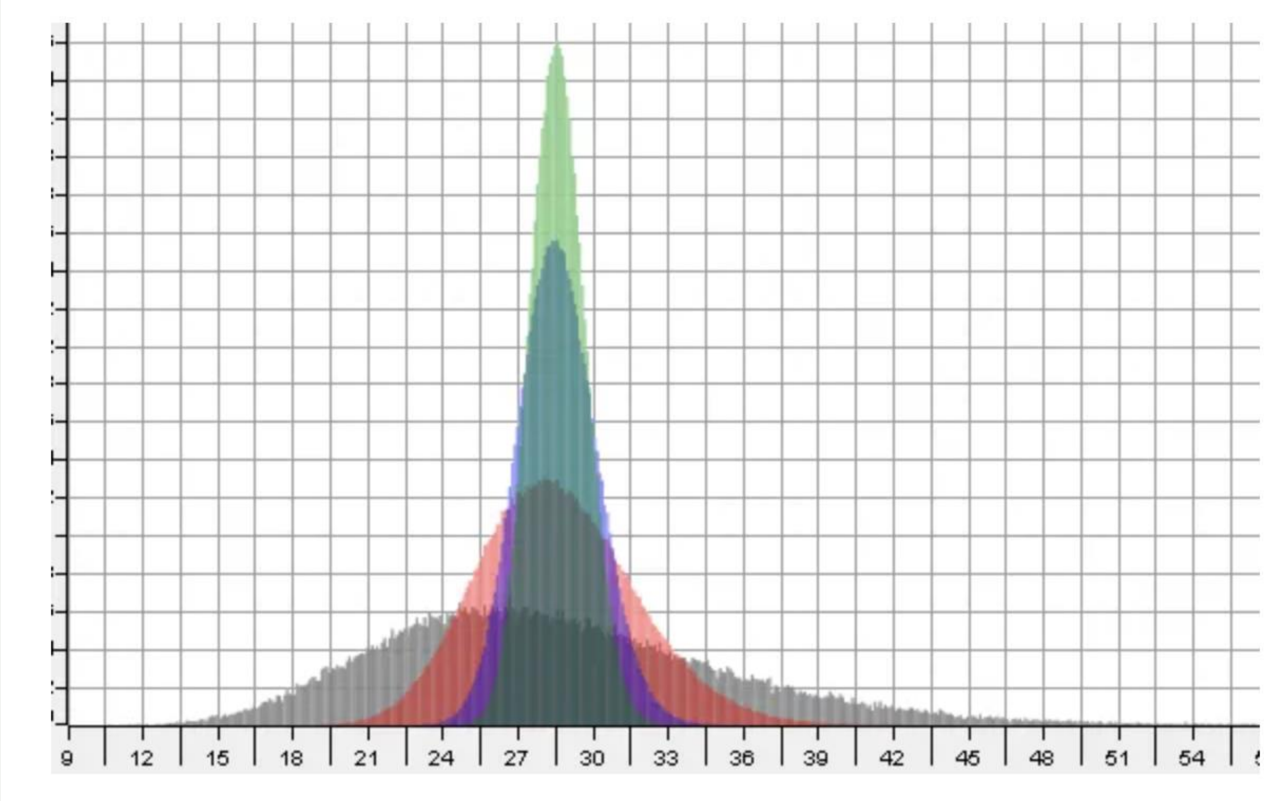
Sampling distribution of…

- the mean, is what CLT deals with
- For each sample, take the mean. Accumulate across say 1000 random draws
- Plot the distribution of these sample means = sampling distribution of the mean

Grey = population
Red = sample n = 5
Blue = sample n = 10
Green = sample n = 20

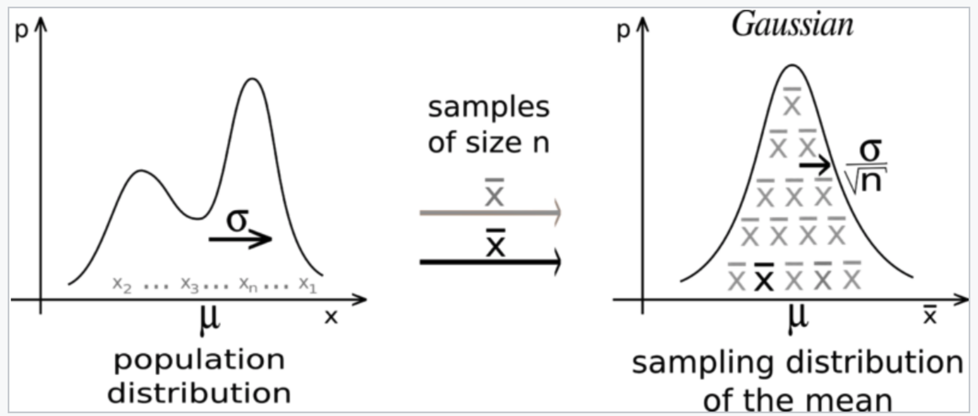
Sample size

- For CLT to work, we need a sufficient sample size when we randomly draw samples **with replacement** from the population. The exact number will depend on the population distribution. Skewed distributions tend to need higher n.
- The sample mean will be equal to the population mean



Lindeberg–Lévy CLT. Suppose $\{X_1, \ldots, X_n\}$ is a sequence of i.i.d. random variables with $\mathbb{E}[X_i] = \mu$ and $\mathrm{Var}[X_i] = \sigma^2 < \infty$. Then as n approaches infinity, the random variables $\sqrt{n}(\bar{X}_n - \mu)$ converge in distribution to a normal $\mathcal{N}(0, \sigma^2)$:[4]

$$\sqrt{n}\left(\bar{X}_n - \mu\right) \xrightarrow{d} \mathcal{N}\left(0, \sigma^2\right).$$



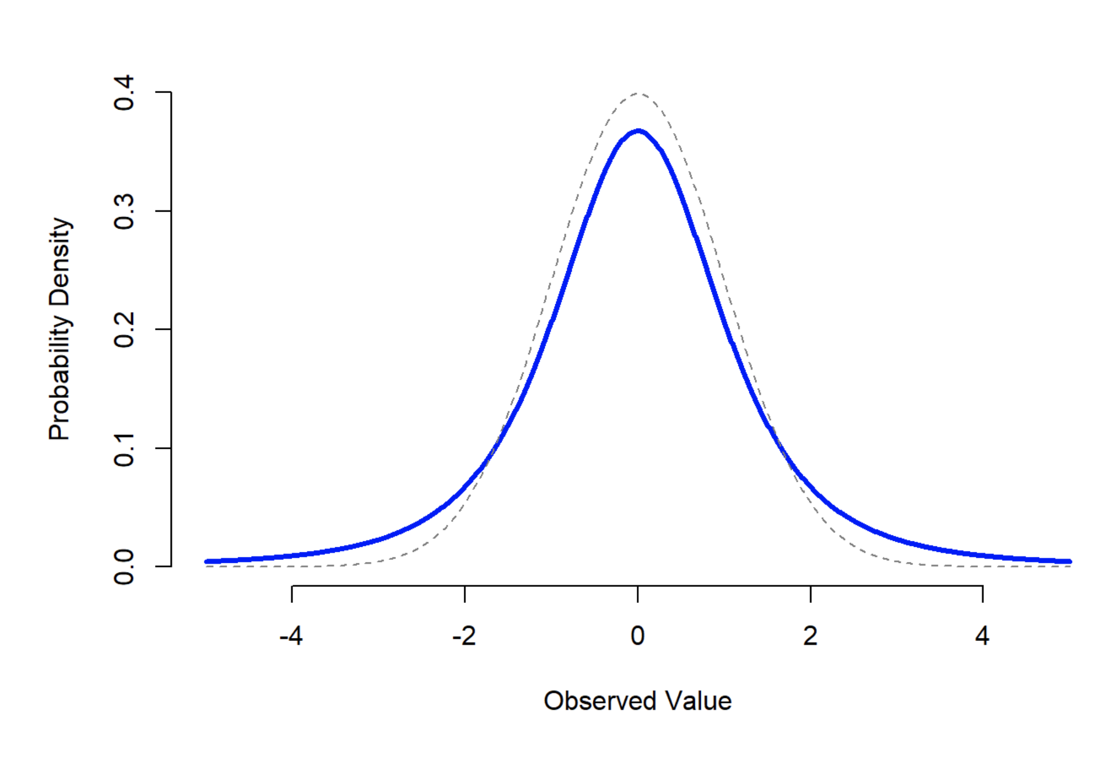
Whatever the form of the population distribution, the sampling distribution tends to a Gaussian, and its dispersion is given by the central limit theorem.[3]

See Wikipedia for an extended introduction to the various forms of the central limit theorem

Why is the central limit theorem important?

- When we test hypotheses about the means of samples (e.g. did healthy adults have a better average performance on my memory task than older adults with MCI?), the tests are often based on the assumption of normality of sampling distributions of the mean.
- CLT says that even if you violate normality assumptions of the variable in the population, as long as you have a sufficiently large sample size, your statistical methods will often be robust to violations of the normality assumptions.

Other distributions: t-distribution



Heavy-tailed
Arises in smaller n situations and when you don't know the population s.d.
As n--> inf, t-distribution begins to look more like a Normal.

T-distributions and k

Degrees of freedom, k, is related to sample size

You can appreciate that as k increases, the shape looks more like a Normal (or the tail gets less heavy).

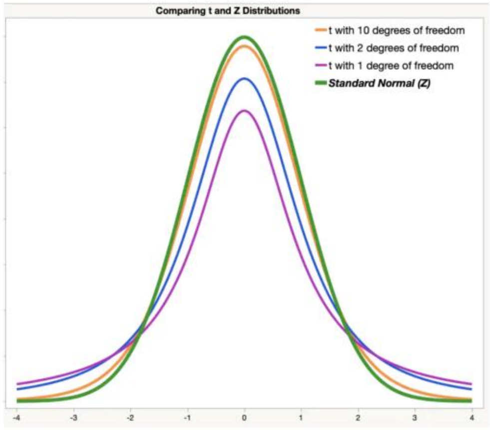


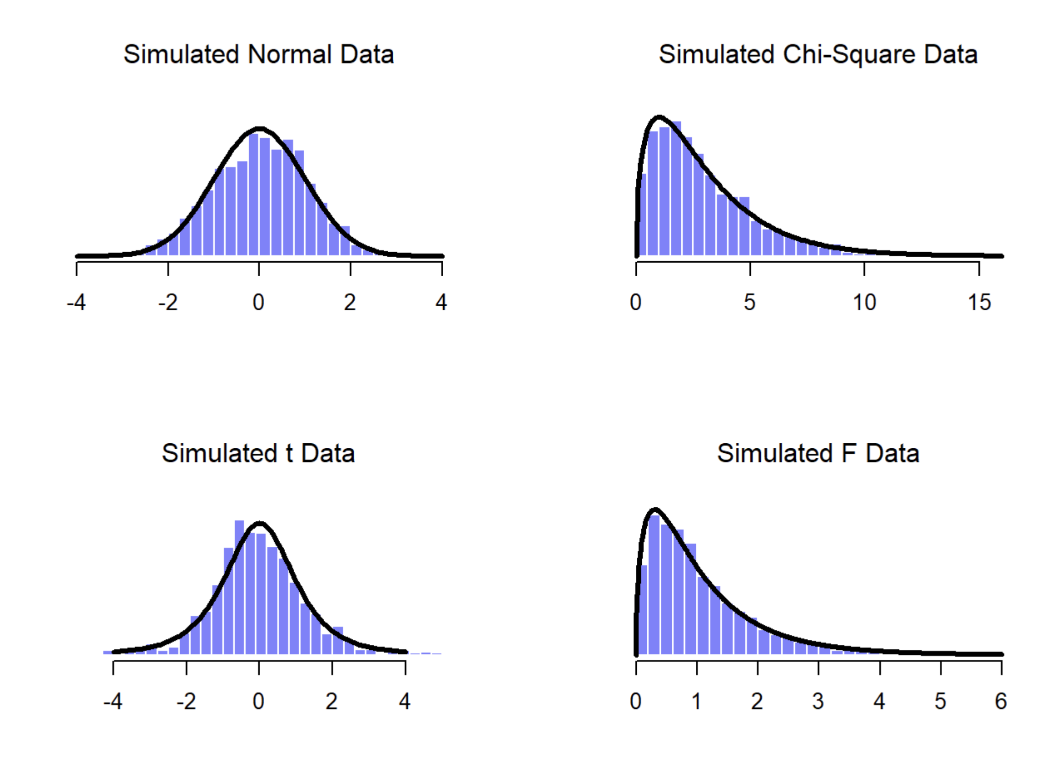
Figure 1: Three t-distributions and a standard normal (z-) distribution.

The use of t-distributions later

Suppose $x_i \sim \mathrm{N}(\mu, \sigma^2)$ and we want to test $H_0 : \mu = \mu_0$ versus $H_1 : \mu \neq \mu_0$.

Assuming we do not know sigma, we will construct a statistic which is where we will encounter the t-distribution to use to construct confidence intervals and p-values to test the above hypothesis

Other distributions



Sum of squares of normally distributed variables: Chi-square

Comparing chi-square distributions: F distributions

Chi-square

- All these other distributions we talk about now are related to the Normal
- chi-square distribution with k degrees of freedom is what you get when you take k normally-distributed variables (with mean 0 and standard deviation 1), square them, and add them up.

```
normal.a <- rnorm( n=1000, mean=0, sd=1 )
normal.b <- rnorm( n=1000 )  # another set of normally distributed data
normal.c <- rnorm( n=1000 )  # and another!
chi.sq.3 <- (normal.a)^2 + (normal.b)^2 + (normal.c)^2
```

R exercises