Topics in Multivariate Analysis

APSTA GE-2004

Lecture 3 - Distributions and CLT

2/8/2022

Outline

Welcome! Today, we'll cover the following:

- Probability distributions
- Sampling distributions
- Sampling and simulation

Reading:

RAOS 3.5-3.6; Ch 4; Ch 5
[Optional]: Modern Dive Ch 7 Sampling
[Optional]: Monte Carlo in R
[Optional]: Discrete and Continuous distributions
[Optional]: Discrete and Continuous distributions

Need to understand probability distributions & simulation

Our language for describing models relies on probability distributions, so let's get familiar with some common distributions.

Similarly, we use sampling and simulation to conduct prior and posterior predictive checks, and to calculate quantities of interest, so let's get familiar with sampling and simulation too.

Probability distributions

Probability versus Inference

Probability: Assumes the data generating process

```
# Probability: assume the probability that a baby is a girl is 48.8%
n_sims <- 1000
n_girls <- rbinom(n = n_sims, size = 400, prob = 0.488)
mean(n_girls) / 400
hist(n_girls)
```

Inference: Learns the data generating process

```
# Inference: learn the probability that a baby is a girl based on data
prop_girls <- n_girls / 400

# Model: fractional logistic regression
fit_frac <- glm(prop_girls ~ 1,
                data = data.frame(prop_girls),
                family = quasibinomial())
tidy(fit_frac)

# extract the Intercept and
# convert logit to a probability (or proportion in this case)
fit_frac %>%
  tidy() %>%
  filter(term == "(Intercept)") %>%
  pull(estimate) %>%
  plogis(.)
```

R's distribution naming conventions

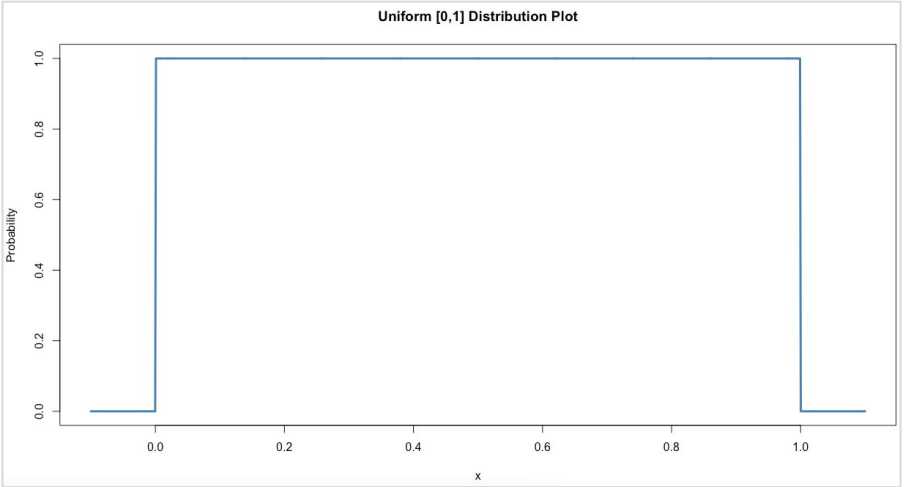
- `dDIST(x, ...)` is the *distribution function (PDF)* that returns the probability of observing the value `x`
- `pDIST(x, ...)` is the *cumulative distribution function (CDF)* that returns the probability of observing a value less than `x`. The flag `lower.tail=F` will cause the function to return the probability of observing a value greater than `x` (the area under the right tail, rather than the left)
- `rDIST(n, ...)` is the random number generation function that returns `n` values drawn from the distribution `DIST`
- `qDIST(p, ...)` is the quantile function that returns the `x` corresponding to the pth percentile of `DIST`. The flag `lower.tail=F` will cause the function to return the `x` that corresponds to the 1 - `p`th percentile of `DIST`

Uniform distribution: on the interval [0, 1]

First principles about the process that generates Y_i:

- Y_i always falls in the “unit” interval
- $Pr(Y \in (a, b)) = Pr(Y \in (c, d))$ if $a < b$, $c < d$, and $b - a = d - c$

- Mean: $(a+b) / 2$
- Variance: $(b-a)^2 / 12$

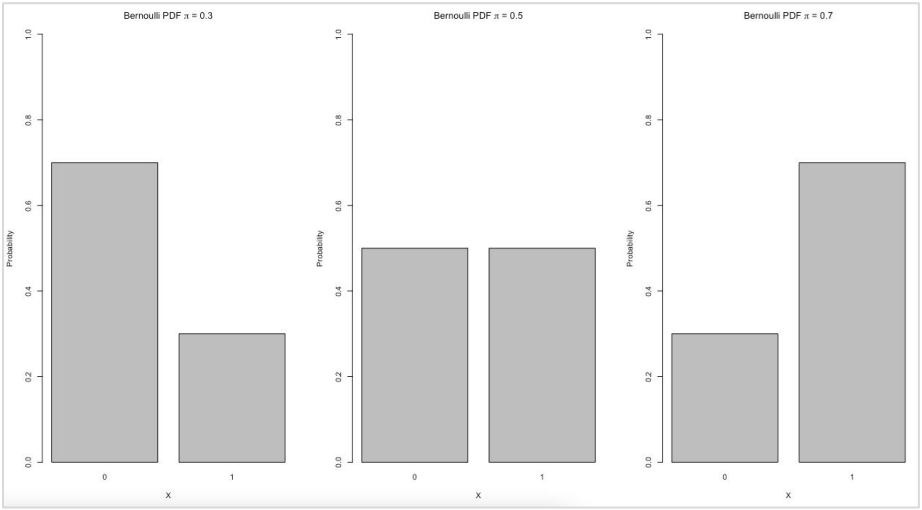



```
> # Area under the curve between 0.25 and 0.75
> punif(0.75) - punif(0.25)
[1] 0.5
>
> # x value corresponding to 75th percentile
> qunif(0.75)
[1] 0.75
>
> # 28 pseudo-random numbers from the uniform distribution
> set.seed(3)
> runif(28)
 [1] 0.16804153 0.80751640 0.38494235 0.32773432 0.60210067 0.60439405 0.12463344 0.29460092 0.57760992 0.63097927 0.51201590 0.50502391 0.53403535 0.55724944
[15] 0.86791949 0.82970869 0.11144915 0.70368836 0.89748826 0.27973255 0.22820188 0.01532989 0.12898156 0.09338193 0.23688501 0.79114741 0.59973157 0.91014771
```

Bernoulli distribution

First principles about the process that generates Y_i:

- Y_i has 2 mutually exclusive outcomes, and
- The 2 outcomes are exhaustive
- The parameter π happens to be interpretable as a probability
  - $P(Y_i=1 \mid \pi_i) = \pi_i \quad , \quad P(Y_i=0 \mid \pi_i) = 1 - \pi_i$
  - $P(Y_i=y \mid \pi_i) = \pi_i^y(1 - \pi_i)^{1-y}$
- Expected value: π
- Variance: π(1 - π)




```
> # Mean, variance, standard deviation
> mean(rbinom(1e5, size = 1, p = 0.3))
[1] 0.29819
> var( rbinom(1e5, size = 1, p = 0.3))
[1] 0.2100821
> sd(  rbinom(1e5, size = 1, p = 0.3))
[1] 0.4592353
>
> # 162 pseudo-random numbers from the uniform distribution
> set.seed(3)
> rbinom(162, size = 1, p = 0.3)
  [1] 0 1 0 0 0 0 0 0 0 0 0 0 0 1 1 0 1 1 0 0 0 0 0 1 0 1 0 1 0 0 0 0 0 0 1 0 0 0 0 1 0 0 0 0 0 0 0 0 1 0 0 1 1 1 1 0 0 0 0 1 0 1 0 1 0 1 0 0 0 1 1 1 1 0 0 0 0 1 1 1
 [82] 0 0 0 1 1 0 0 0 1 0 0 0 1 1 0 0 0 0 0 1 0 0 1 0 0 0 0 0 0 0 0 0 0 1 1 1 0 0 0 1 1 1 0 0 0 0 0 1 0 0 0 0 0 1 0 0 0 0 0 0 0 0 1 1 0 0 1 0 0 1 0 1 0 0 0 0 0 0 0 1 0
```

Binomial distribution

First principles about the process that generates Y_i:

- N **iid** Bernoulli trials, y_1, ..., y_N
- The trials are **independent**
- The trials are **identically distributed**
- We observe $Y = \Sigma\, y_i$

Density:

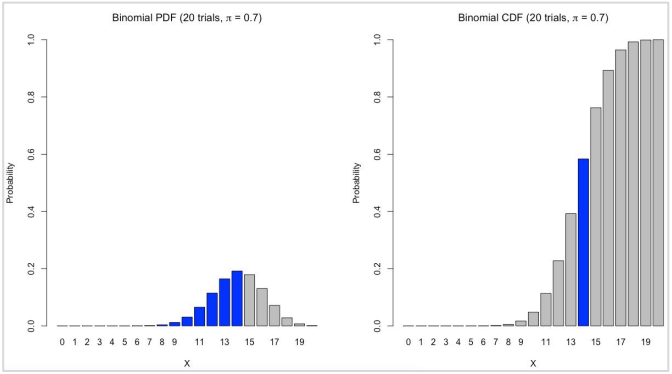
$$P(Y = y \mid \pi) = (N \text{ choose } y)\, \pi^y (1 - \pi)^{1-y}$$

Explanation:

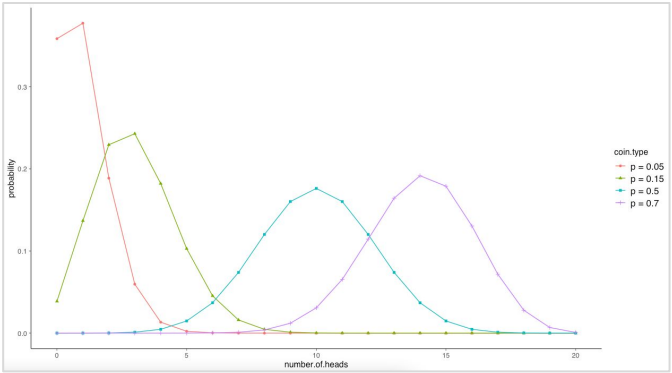
- (N choose y) because (1 0 1) and (1 1 0) are both y = 2
- π^y because y successes with π probability each (product due to independence)

- Mean: Nπ
- Variance: Nπ(1-π)

1 process: 20 trials (70% chance of success on each trial)



4 processes: 20 trials (each process has a different chance of success on each trial)



Normal distribution

First principles about the process that generates Y_i:

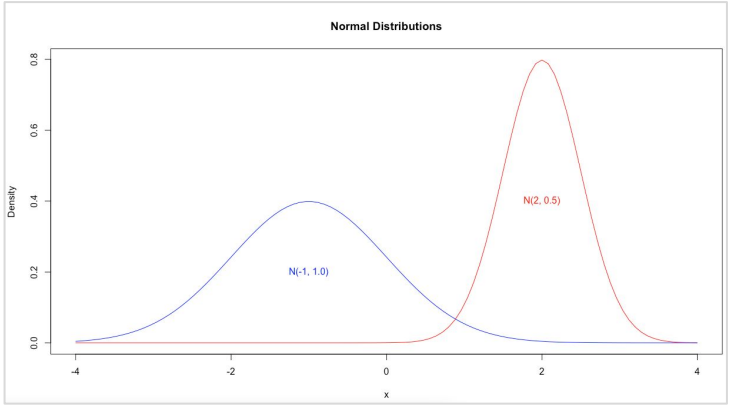
- Many different first principles
  - The Central Limit Theorem
  - Normal by addition
  - Normal by multiplication
  - Normal by log-multiplication
- The univariate normal density (with mean μ_i, variance σ^2):

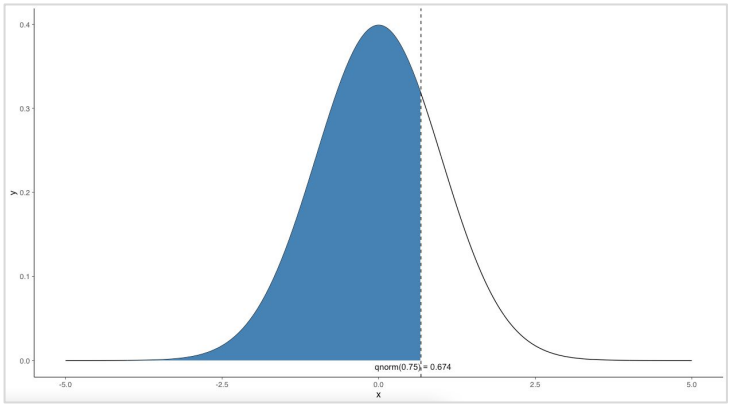
  $$N(y_i | \mu_i, \sigma^2) = (2\pi\sigma^2)^{-1/2} \exp(-(y_i - \mu_i)^2 / 2\sigma^2)$$

- Regression notation:

  $Yi \sim N(\mu_i, \sigma^2)$ stochastic

  $\mu_i = x_i\beta$ systematic



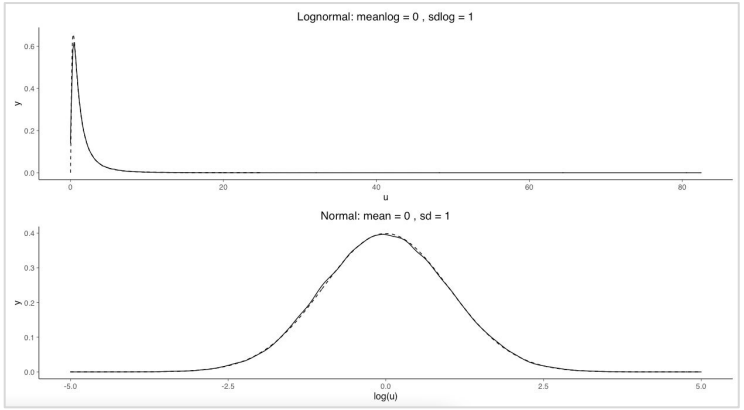


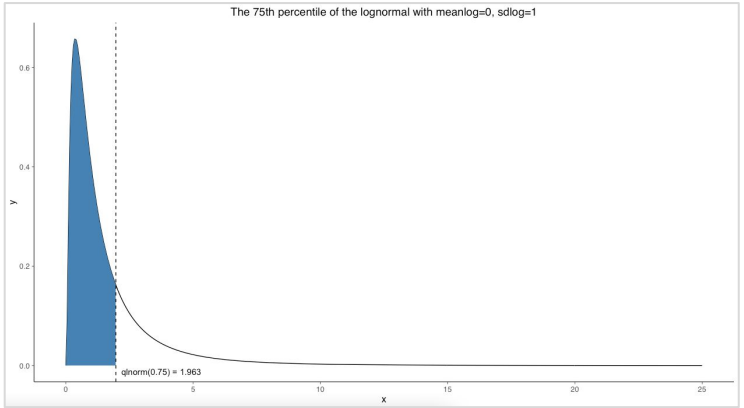

```
> # 50% of the observations will be less than the mean
> pnorm(0)
[1] 0.5
>
> # About 2.3% of the observations are more than 2
> # standard deviations below the mean
> pnorm(-2)
[1] 0.02275013
>
> # About 95.4% of the observations are within 2
> # standard deviations from the mean
> pnorm(2) - pnorm(-2)
[1] 0.9544997
```

Lognormal distribution

First principles about the process that generates Y_i :

- A random variable whose *natural log* is normally distributed
- Defined over all non-negative real numbers:
  - A distribution of highly skewed positive data
  - Mean is generally much higher than the median
- Useful if variations in the data are expressed naturally as percentages or relative differences, rather than absolute differences, e.g.:
  - Incomes (5% increase in salary; not $5,000/yr for everyone)
  - Revenue (project to within 10%; not within +/- $1,000)
  - Sales
  - Stock prices






```
> # the 50th percentile (median) of the lognormal
> # with meanlog=0 and sdlog=1
> qlnorm(0.5)
[1] 1
>
> # the probability of seeing a value x less than 1
> plnorm(1)
[1] 0.5
>
> # the probability of seeing a value x less than 10
> plnorm(10)
[1] 0.9893489
```

Exponential distribution

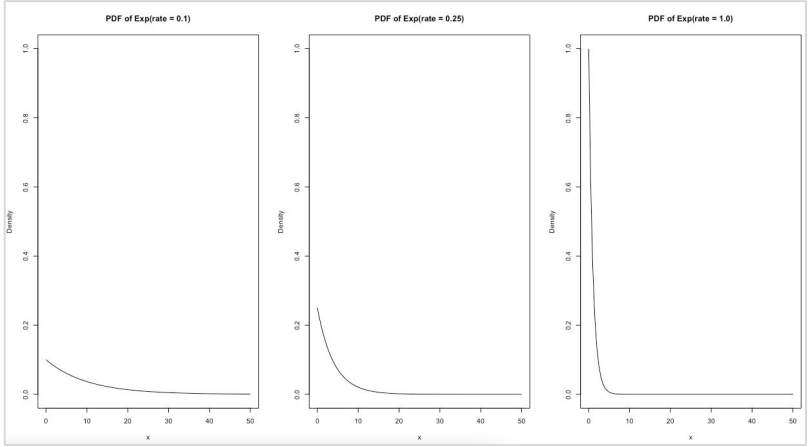
First principles about the process that generates Y_i :

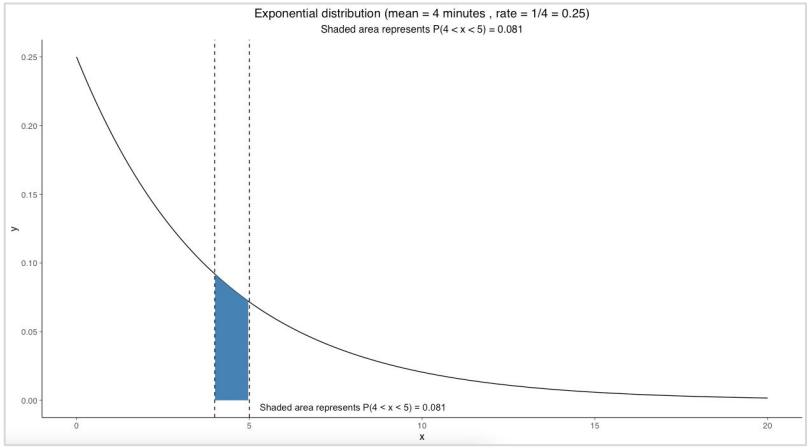
- Often represents the arrival time of a randomly recurring independent event sequence:
  - Amount of time (beginning now) until an earthquake
  - Amount of time a telephone call lasts
  - Amount of time a product lasts
  - Also, how much money people spend in one trip to store

- If μ is the *mean* waiting time for the next event recurrence, its probability density function is:
  - $f(x) = 1/\mu\, e^{-x/\mu}$ when $x \geq 0$; else 0 when $x < 0$

- Values for an exponential RV occur in the following way:
  - There are fewer large values and more small values
  - For example, the amount of money customers spend in one trip to the supermarket. There are more people who spend small amounts of money and fewer people who spend large amounts

```
> # the 50th percentile (median) of the exponential
> # with rate=1
> qexp(0.5, rate = 1)
[1] 0.6931472
> pexp(qexp(0.5, rate = 1), rate = 1)
[1] 0.5
>
> # the probability of seeing a value x less than 2
> pexp(2, rate = 1)
[1] 0.8646647
>
> # the probability of seeing a value x less than 5
> pexp(5, rate = 1)
[1] 0.9932621
```





Poisson distribution

First principles about the process that generates Y_i :

- Often represents the number of times an event occurs (count data) in a *fixed* interval of time or space:
  - Number of tropical cyclones crossing coast during season
  - Number of people who buy a product in a week
  - Number of vehicle crashes in a year
  - Number of occupational injuries in a month

- If *Y* is the number of occurrences, its probability density function is:

  $f(y) = \lambda^y e^{-\lambda} / y! \quad ; y = 0, 1, 2,\ldots$

  where **y** is a non-negative **integer** and **λ** is a non-negative **real number** (λ is often described as a rate (e.g. crashes per 100,000 km per year), and the time scale should be included in the definition)

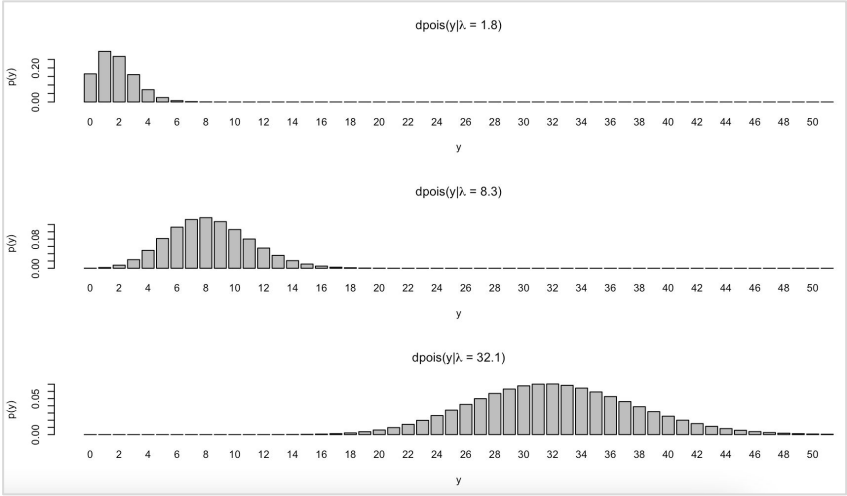
- Importantly, the **λ is both the mean and variance** of the Poisson distribution (the standard deviation is **√λ**)

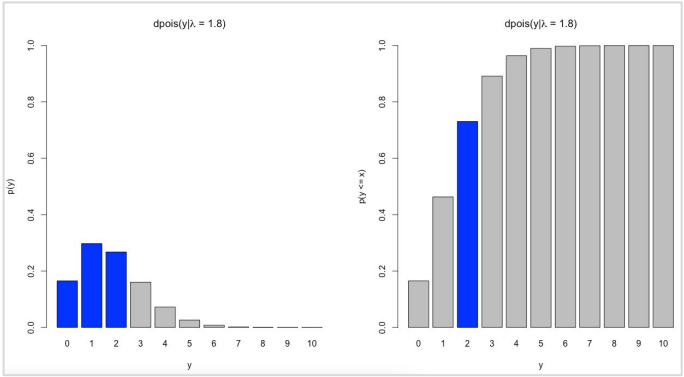
- Generally, the rate is specified in terms of "exposure"; for instance, customers entering a store are "exposed" to the opportunity to buy

```
# the 50th percentile (median) for Poisson(lambda = 1.8)
qpois(0.5, lambda = 1.8)

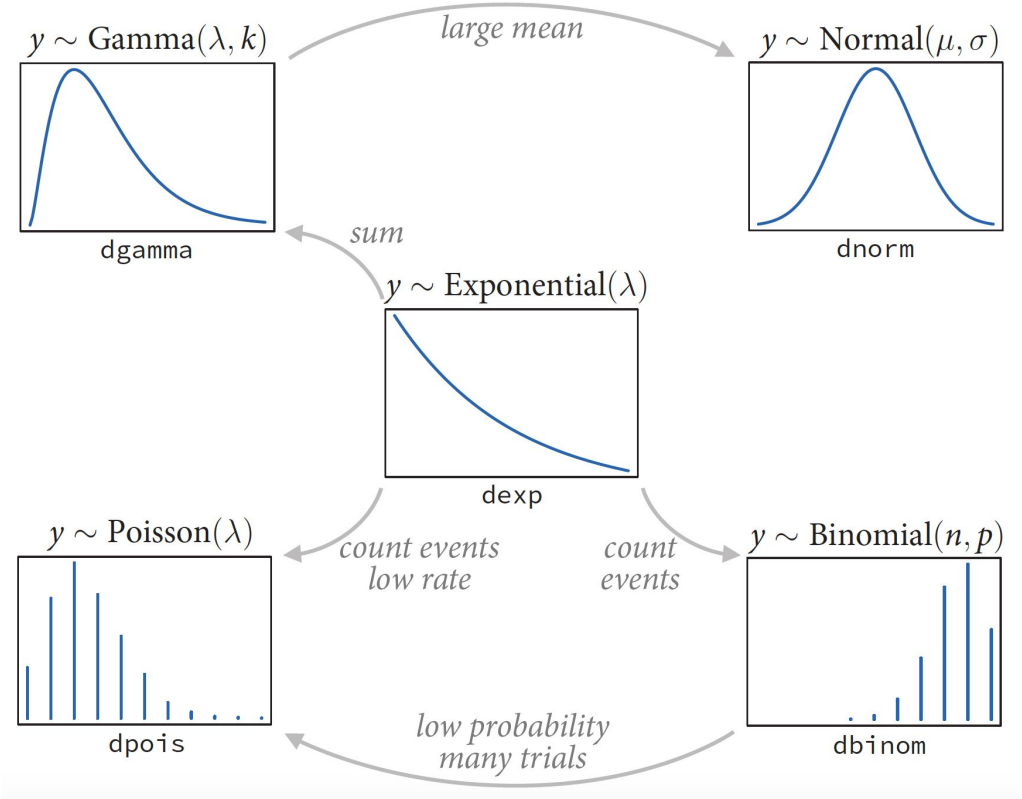
# the probability of 2 occurrences for Poisson(lambda = 1.8)
dpois(2, lambda = 1.8)

# the probability of at most 2 occurrences for Poisson(lambda = 1.8)
ppois(2, lambda = 1.8)
```





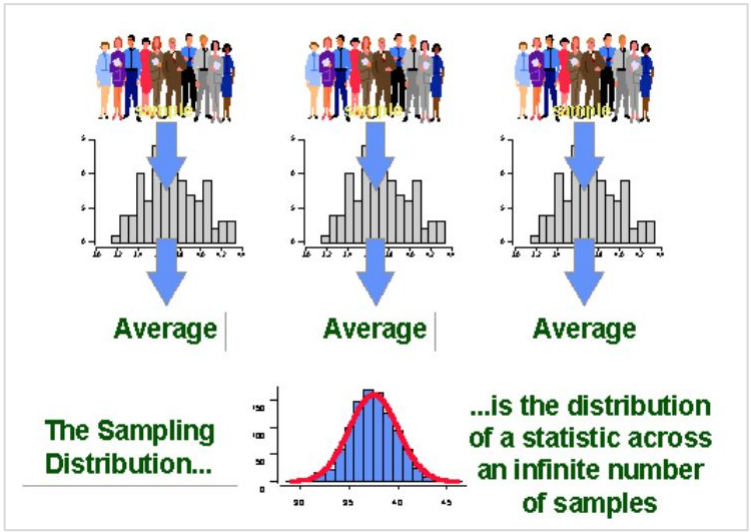
Relationships between exponential family of distributions

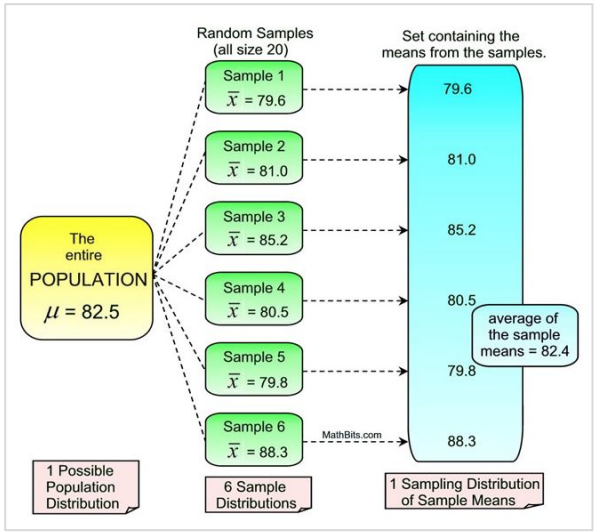


Sampling distributions

Sampling distribution

The *sampling distribution* is the set of possible datasets that could have been observed if the data collection process had been re-done, along with the probabilities of these possible values





Sampling distribution of a count

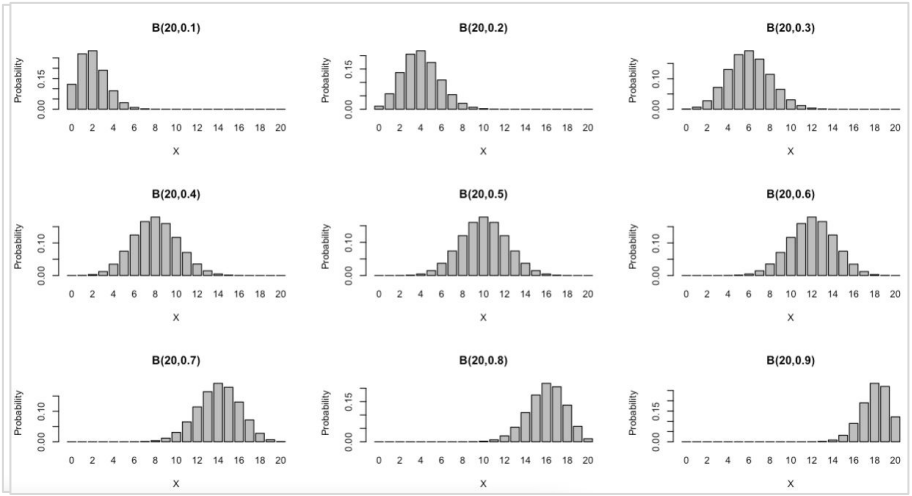
The Binomial setting:

- There are a fixed number ***n*** of observations
- The ***n*** observations are all independent
- Each observation falls into one of just two categories, which for convenience we call “success” and “failure”
- The probability of a “success”, call it ***p***, is the same for each observation

Binomial distribution: *B*(*n*,*p*)

- The distribution of the count **X** of successes in the binomial setting is called the *binomial distribution* with parameters ***n*** and ***p***
- ***n*** is the number of observations, and ***p*** is the probability of a success on any one observation
- The possible values of **X** are the whole numbers from **0** to ***n***

20 observations ; p varies from 0.1 to 0.9



Binomial mean and standard deviation

If a count **X** is $B(n,p)$, what are the mean μ_X and the standard deviation σ_X?

Mean (μ_X):

Let the random variable $S_i = 1$ if "success" and $S_i = 0$ if "failure"; then the mean of each S_i is:

$\mu_S = (1)(p) + (0)(1 - p) = p$

Because each S_i is 1 for "success" and 0 for "failure", the total number of "successes" X is the sum of the S_i's:

$X = S_1 + S_2 + \ldots + S_n$

By the *addition rule for means*, the mean of **X** is the sum of the means of the S_i's

$\mu_X = \mu_{S1} + \mu_{S2} + \ldots + \mu_{Sn}$

$= n\mu_S$

$\mu_X = np$

Standard deviation (σ_X):

Let the random variable $S_i = 1$ if "success" and $S_i = 0$ if "failure"; then the *variance* of each S_i is:

$\sigma^2_S = (1-p)^2(p) + (0 - p)^2(1-p)$

$= (p)(1-2p+p^2) + (p^2)(1-p)$

$= p^2 + p - 2p^2$

$= p - p^2$

$= p(1-p)$

By the *addition rule for variances*, the variance of **X** is *n* times the variance of a single S, so:

$\sigma^2_X = np(1-p)$

The standard deviation of **X** is the square root of the variance:

$\sigma_X = \sqrt{np(1-p)}$

Sample proportion mean and standard deviation

We often want to estimate the proportion p of "successes" in a population. Our estimator is the sample proportion of "successes":

p_hat = count of successes in sample / size of sample
= X / n

Be sure to distinguish between the *proportion* ***p_hat*** and the *count* ***X***. The count takes whole-number values between 0 and n, but a proportion is always a number between 0 and 1.

In the binomial setting, the *count* ***X*** has a binomial distribution. The *proportion* ***p_hat*** does not have a binomial distribution (*it's approx. normal when n is large*); however, we can do probability calculations about ***p_hat*** by restating them in terms of the *count* ***X*** and using binomial methods.

We can obtain the mean and standard deviation of the sample proportion from the mean and standard deviation of the sample count using the rules for the mean and variance of a constant times a random variable:

Mean: $\mu_{p\_hat} = p$

Standard deviation: $\sigma_{p\_hat} = \sqrt{p(1-p)/n} = \sqrt{p(1-p)} / \sqrt{n}$

The √n in the denominator means that the sample size must be multiplied by 4 if we wish to divide the standard deviation in half

Normal approximation for counts and proportions

Let ***X*** be the count of “successes” in the sample and ***p_hat = X/n*** the sample proportion of “successes”.

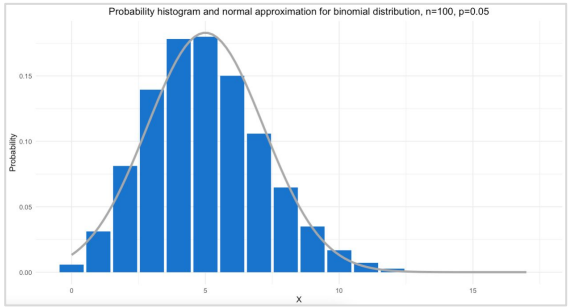
In large samples, both the *count X* and the *proportion p_hat* are approximately normal:

count X is approximately ***N(np , √np(1–p))***

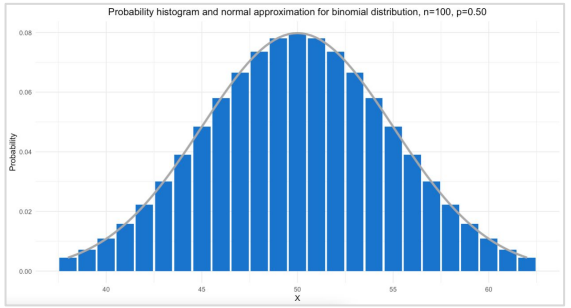
proportion p_hat is approximately ***N(p , √p(1–p)/n)***

The accuracy of the normal approximations improves as the sample size *n* increases, and they are most accurate for any fixed *n* when *p* is close to ½, and least accurate when *p* is near 0 or 1

Less accurate normal approximation



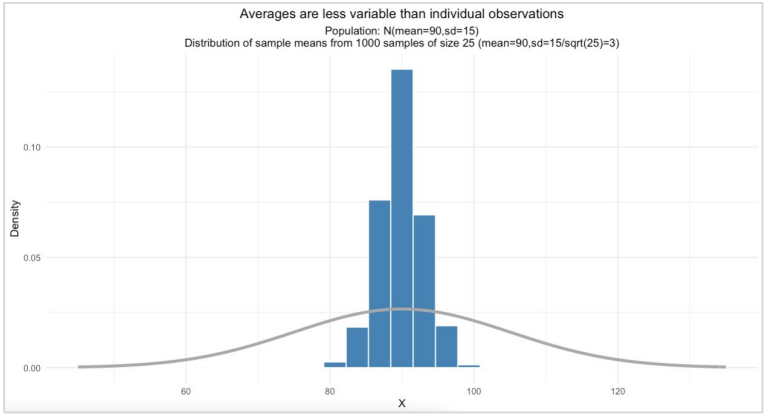
More accurate normal approximation

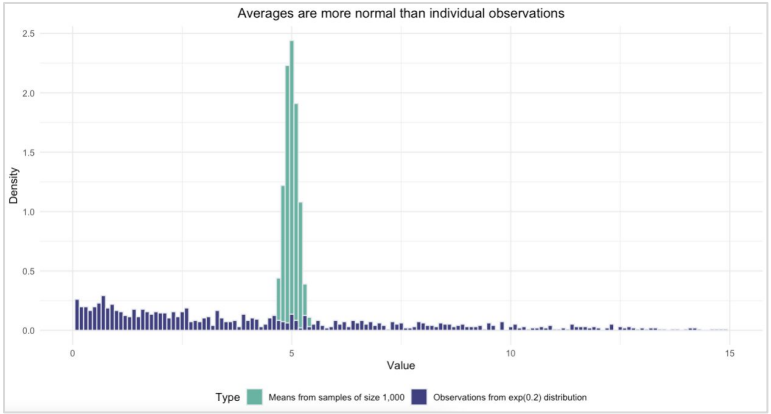


Sampling distribution of a sample mean

Two facts that contribute to the popularity of sample means:

- Averages are *less variable* than individual observations
- Averages are *more normal* than individual observations





Mean and standard deviation of sample mean

Sampling distribution of a sample mean (***x_bar***):

If a population has the N(μ, σ) distribution, then
the ***sample mean x_bar*** of ***n*** independent observations has the **N(μ, σ/√n)** distribution

Mean (μ_{x_bar}):

Select a random sample of size *n*, and measure a variable X on each individual in the sample. Then the *sample mean* is:

$\mu_{x\_bar} = 1/n * (X_1 + X_2 + \ldots + X_n)$

If the population has mean μ, then μ is the mean of each observation X_i. By the *addition rule for means*, the mean of the sample is:

$\boldsymbol{\mu_{x\_bar}} = 1/n * (\mu_{x1} + \mu_{x2} + \ldots + \mu_{xn})$

$= 1/n * (\mu + \mu + \ldots + \mu)$

$\boldsymbol{\mu_{x\_bar} = \mu}$

That is, the mean of x_bar is the same as the mean of the population

Standard deviation (σ_{x_bar}):

Select a random sample of size *n*, and measure a variable X on each individual in the sample. By the *addition rule for variances*, the *sample variance* is:

$\sigma^2_{x\_bar} = (1/n)^2 * (\sigma^2_{x1} + \sigma^2_{x2} + \ldots + \sigma^2_{xn})$

$= (1/n)^2 * (\sigma^2 + \sigma^2 + \ldots + \sigma^2)$

$= \sigma^2/n$

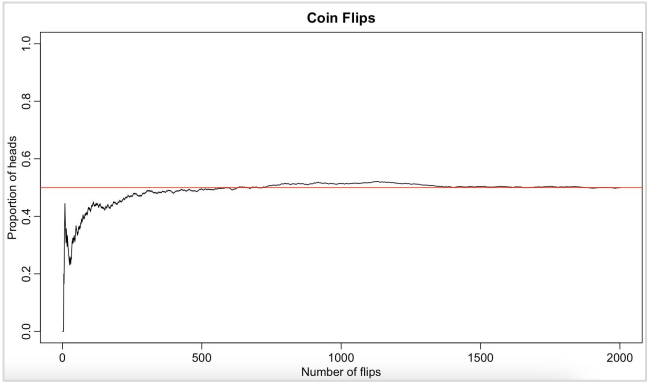
The standard deviation of x_bar is the square root of the variance:

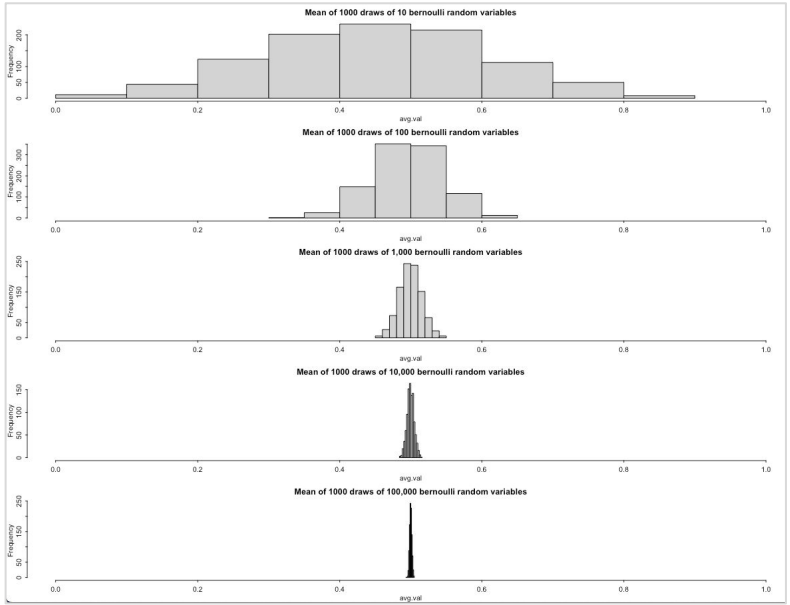
$\boldsymbol{\sigma_{x\_bar} = \sigma / \sqrt{n}}$

The variability of the sampling distribution of a sample mean decreases as the sample size grows; given the √n, it decreases in proportion to the square root of the sample size

Law of Large Numbers

- The LLN is a theorem that describes the result of performing the same experiment a large number of times
- According to the law, the *average* of the results obtained from a large number of trials should be close to the *expected value* and will tend to become closer to the expected value as more trials are performed
- The LLN is important because it guarantees stable long-term results for the averages of some random events





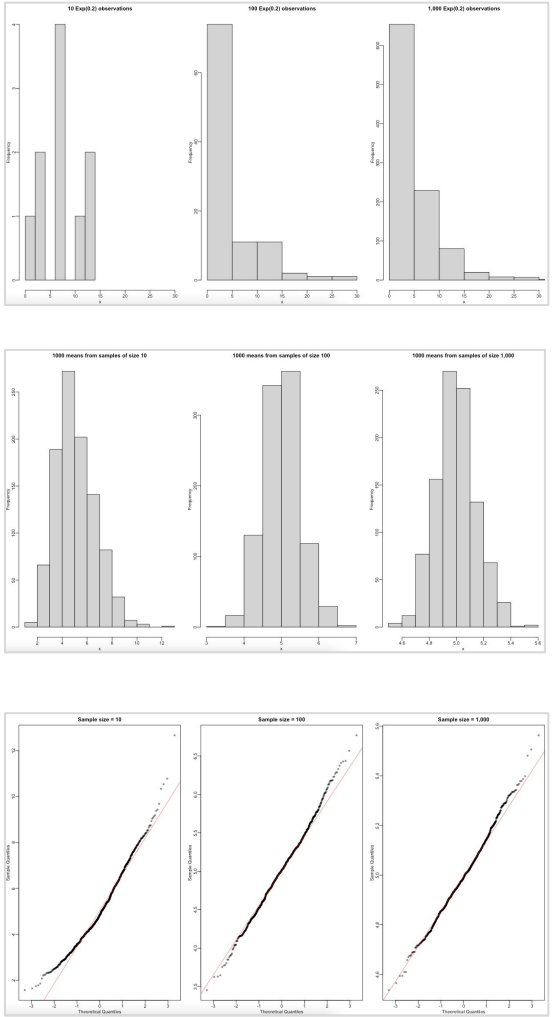
Central Limit Theorem

- The CLT establishes that, in many situations, when independent random variables are summed up, their properly normalized sum tends toward a normal distribution (informally a bell curve) even if the original variables themselves are not normally distributed

- The theorem implies that probabilistic and statistical methods that work for normal distributions can be applicable to many problems involving other types of distributions

- For example, suppose a sample is obtained containing many observations, each observation being randomly generated in a way that does not depend on the values of the other observations, and that the arithmetic mean of the observed values is computed.

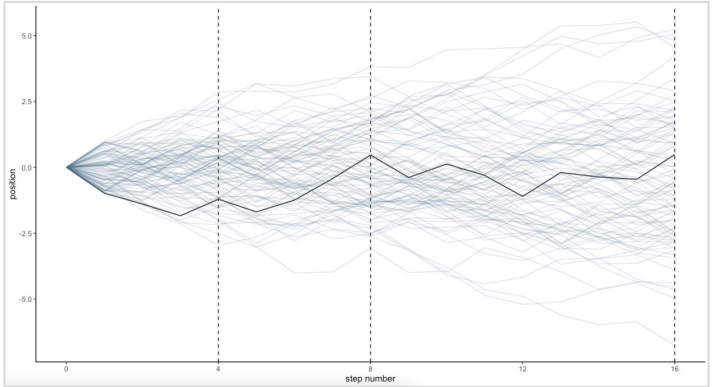
  If this procedure is performed many times, the CLT says that the probability distribution of the average will closely approximate a normal distribution

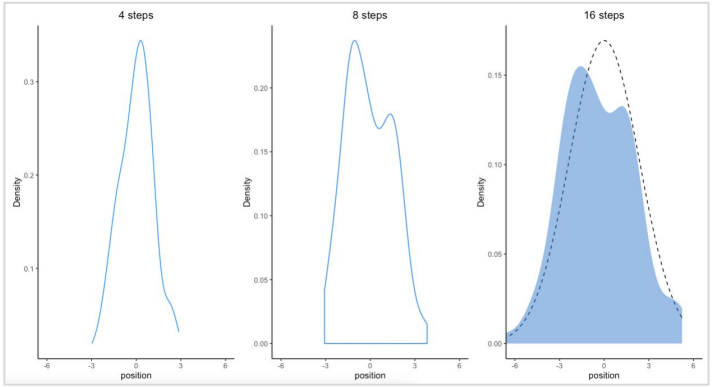


Why are normal distributions normal?

- Normal by addition
  - Any process that add together random values from the same distribution converges to a normal
  - ```
    # Normal by addition
    pos <- replicate( 1000 , sum( runif(16 , -1 , 1) ) )
    hist(pos)
    plot(density(pos), xlab = "", main = "Normal by addition")
    ```
- Normal by multiplication
  - Any process that multiplies small deviations together tends to converge to a normal because multiplying small numbers is approximately the same as addition
  - ```
    # Normal by multiplication
    growth <- replicate( 10000 , prod( 1 + runif(12 , 0 , 0.1) ) )
    hist(growth)
    plot(density(growth), xlab = "", main = "Normal by multiplication")
    ```
- Normal by log-multiplication
  - Any process that multiplies large deviations together tends to converge to a normal when measured on the log scale because adding logs is equivalent to multiplying the original numbers
  - ```
    # Normal by log-multiplication
    log.big <- replicate( 10000 , log( prod( 1 + runif(12 , 0 , 0.5) ) ) )
    hist(log.big)
    plot(density(log.big), xlab = "", main = "Normal by log-multiplication")
    ```

Since measurement scales are arbitrary, all of these methods are legitimate




    ```

Rules for means

Rule 1: If **X** and **Y** are random variables (rv), then

$$\mu_{\mathbf{X+Y}} = \mu_{\mathbf{X}} + \mu_{\mathbf{Y}}$$

Example:
The number of dimples on a fridge is a rv **X** that takes values 0, 1, 2, and so on. **X** varies from fridge to fridge. The mean number of dimples is $\mu_{\mathbf{X}} = 0.7$

Similarly, the number of paint sags is a second rv **Y** that takes values 0, 1, 2, and so on. **Y** varies from fridge to fridge. The mean number of paint sags is $\mu_{\mathbf{Y}} = 1.4$

The total number of both dimples and sags is the sum **X** + **Y**. This sum is a rv that varies from fridge to fridge. Its mean $\mu_{\mathbf{X+Y}}$ is the average number of dimples and sags together: *it is the sum of the individual means $\mu_{\boldsymbol{X}}$ and $\mu_{\mathbf{Y}}$*

Rule 2: If **X** is a random variable and ***a*** and ***b*** are fixed numbers, then

$$\mu_{a+b\mathbf{X}} = a + b\mu_{\mathbf{X}}$$

Example:
Suppose **X** is the width in centimeters of a flower chosen from a tree and that the mean width is $\mu_{\mathbf{X}} = 2.2$ centimeters

If we decide to measure in millimeters, we multiply every value of **X** by 10 because there are 10 millimeters in a centimeter. Then, *just as we multiply every value of **X** by 10, we also multiply the mean by 10*. That is, the mean $\mu_{\mathbf{10X}}$ of **10X** is $10\mu_{\mathbf{X}} = 2.2 \times 10 = 22$ millimeters

Similarly, *if we add the same fixed number to every value of a rv **X**, we add that same number to the mean*

Rules for variances

Rule 1: If X and Y are independent random variables (rv), then

$$\sigma^2_{\mathbf{X+Y}} = \sigma^2_{\mathbf{X}} + \sigma^2_{\mathbf{Y}}$$

$$\sigma^2_{\mathbf{X-Y}} = \sigma^2_{\mathbf{X}} + \sigma^2_{\mathbf{Y}}$$

Explanation:

Because the square of –1 is 1, the additional rule says that the variance of a difference is the ***sum*** of the variances. The difference **X** – **Y** is more variable than either **X** or **Y** alone because variations in both **X** and **Y** contribute to variation in their difference

The addition rule for variances implies that standard deviations do ***not*** generally add. For example, the standard deviations of 2X and –2X are both equal to $2\sigma_{\mathbf{X}}$ because this is the square root of the variance $4\sigma^2_{\mathbf{X}}$

Rule 2: If X is a random variable and ***a*** and ***b*** are fixed numbers, then

$$\sigma^2_{a+b\mathbf{X}} = b^2\sigma^2_{\mathbf{X}}$$

Explanation:

Because a variance is the average of ***squared*** deviations from the mean, multiplying **X** by a constant *b* multiplies $\sigma^2_{\mathbf{X}}$ by the ***square*** of the constant

Adding a constant *a* to a rv changes its mean but does ***not*** change its variability. The variance of **X** + *a* is the same as the variance of **X**

Sampling and Simulation

The importance of simulation in applied statistics

In statistics, we use probabilistic models to represent variation in the real world. Since probability and randomness are hard to understand, we can use simulations to train our intuition.

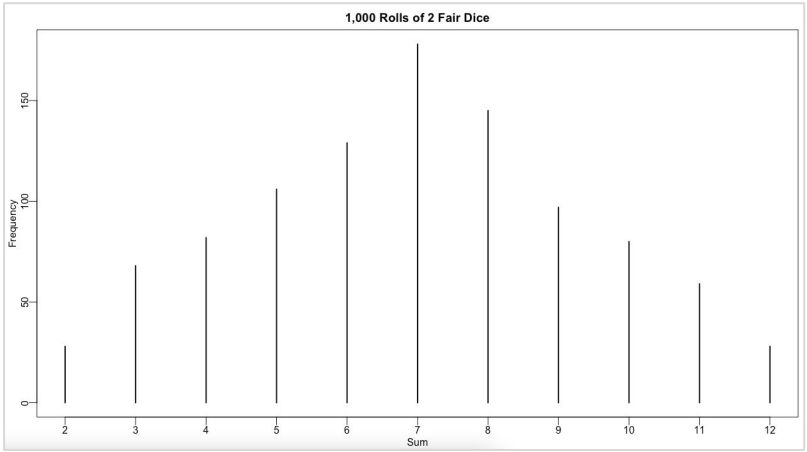
We can use simulations to approximate sampling distributions of data, and therefore to approximate sampling distributions of estimates.

Statistical regression models produce probabilistic predictions; they are not deterministic. Simulations allow us to represent uncertainty in regression predictions and parameter estimates.

Sampling “marbles from a bowl”

- The R command `sample` simulates drawing marbles from a bowl
- Many random experiments can be reduced to thinking about a bowl containing different kinds of marbles, so `sample` is a general command
- The function `sample` takes three arguments:
  - `x` is a vector containing the “marbles”
  - `size` tells R how many marbles we want to draw
  - `replace` is set to `TRUE` or `FALSE` depending on whether we want to sample *with* or *without* replacement

```
> marbles = c('red', 'blue', 'green')
> sample(x = marbles, size = 2, replace = FALSE)
[1] "blue"  "green"
> sample(x = 1:6, size = 1, replace = TRUE)
[1] 1
>
> sample(x = 1:6, size = 2, replace = TRUE)
[1] 5 1
>
> sum(sample(x = 1:6, size = 2, replace = TRUE))
[1] 6
> my.dice.sum <- function(n.dice, n.sides){
+   dice <- sample(x = 1:n.sides, size = n.dice, replace = TRUE)
+   return(sum(dice))
+ }
> sims <- replicate(1e3, my.dice.sum(n.dice = 2, n.sides = 6))
> table(sims)
sims
  2   3   4   5   6   7   8   9  10  11  12
 28  68  82 106 129 178 145  97  80  59  28
> table(sims) / length(sims)
sims
    2     3     4     5     6     7     8     9    10    11    12
0.028 0.068 0.082 0.106 0.129 0.178 0.145 0.097 0.080 0.059 0.028
> plot(table(sims), xlab = "Sum", ylab = "Frequency", main = "1,000 Rolls of 2 Fair Dice")
```



Sampling from a discrete probability distribution

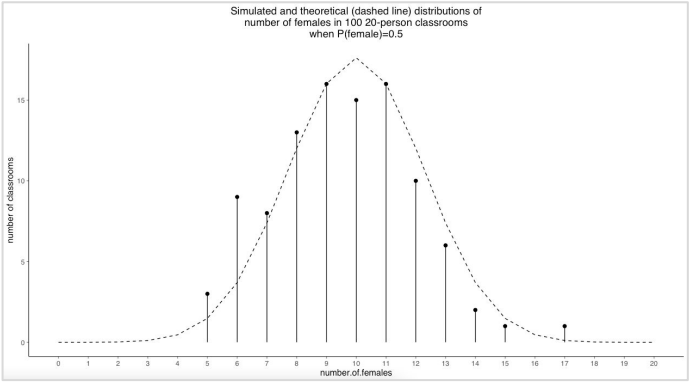
Suppose there is a large student population that is 50% female

If students are assigned to classrooms at random, and you visit 100 classrooms with 20 students each, how many females do you expect to see in each classroom?

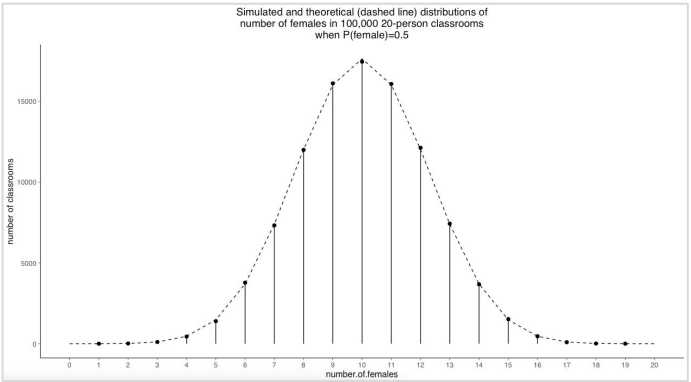
Sampling from a discrete probability distribution

Theoretically, you would expect to see the highest number of classrooms with 10 females, the next highest number of classrooms with 9 or 11 females, the next highest number of classrooms with 8 or 12 females, and so on… (dashed line)

Given that we simulated only 100 classrooms, our simulated distribution doesn't approximate the theoretical distribution very well.



Instead, if we simulate 100,000 classrooms, then our simulated distribution approximates the theoretical distribution more closely.



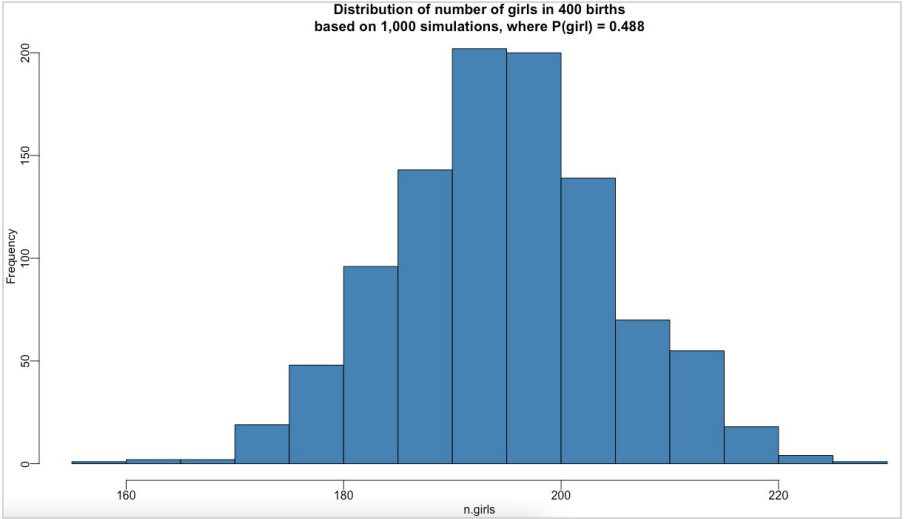
Sampling from a discrete probability distribution

The probability that a baby is a girl or boy is 48.8% or 51.2%, respectively.

Suppose that 400 babies are born in a hospital in a given year, what is the distribution of the number of girls you expect to see?

Sampling from a discrete probability distribution

```r
n.sims <- 1000
n.girls <- rep(NA, n.sims)
for (s in 1:n.sims){ n.girls[s] <- rbinom(n = 1, size = 400, prob = 0.488) }
hist(n.girls, col = "steelblue", main = "")
```



Sampling from a continuous probability distribution

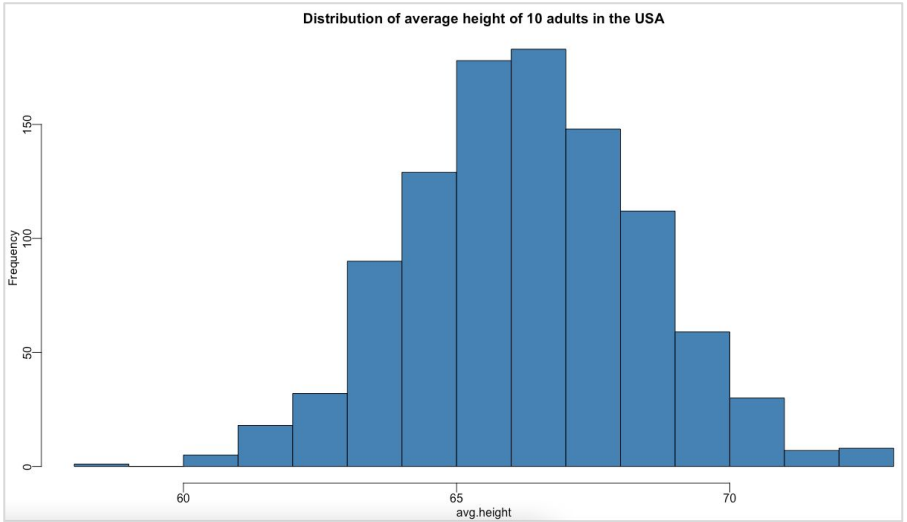
In the United States, 52% of adults are women and 48% are men.

The heights of the men are approximately normally distributed with mean 69.1 inches and sd 2.9 inches; women with mean 63.7 and sd 2.7.

Suppose we select 10 adults at random, what is the distribution of the average height we expect to see?

Sampling from a continuous probability distribution

```r
n.sims <- 1000
avg.height <- rep(NA, n.sims)
max.height <- rep(NA, n.sims)
for (s in 1:n.sims){
  sex <- rbinom(n = 10, size = 1, prob = 0.52)
  height <- ifelse(sex==0, rnorm(n = 1, mean = 69.1, sd = 2.9), rnorm(n = 1, mean = 63.7, sd = 2.7))
  avg.height[s] <- mean(height)
  max.height[s] <-  max(height)
}
hist(avg.height, col = "steelblue", main = "Distribution of average height of 10 adults in the USA")
```



Monte Hall's Let's Make a Deal

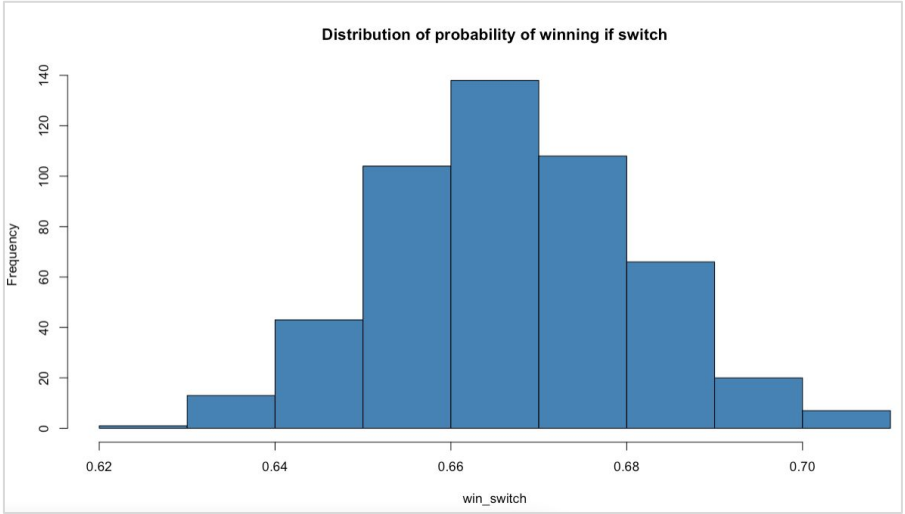
Choose 1 of 3 doors (two with goats; one with a sports car...you want the sports car!).

Monte peeks at the 2 unchosen doors and opens the one (or one of the two) with a goat and asks if you'll switch to the remaining door?

Should you switch?

Monte Hall's Let's Make a Deal

```
> sims <- 1000
> win_no_switch <- 0
> win_switch <- 0
> doors <- c(1, 2, 3)
>
> for (i in 1:sims) {
+   win_door <- sample(x = doors, size = 1)
+   choice   <- sample(x = doors, size = 1)
+   if (win_door == choice)
+       win_no_switch <- win_no_switch + 1
+   doors_remaining <- doors[doors != choice]
+   if( any(doors_remaining == win_door) )
+       win_switch <- win_switch + 1
+ }
>
> cat("Prob(Car | no switch) = " , win_no_switch / sims , "\n")
Prob(Car | no switch) =  0.334
> cat("Prob(Car | switch) = " , win_switch / sims , "\n")
Prob(Car | switch) =  0.666
```



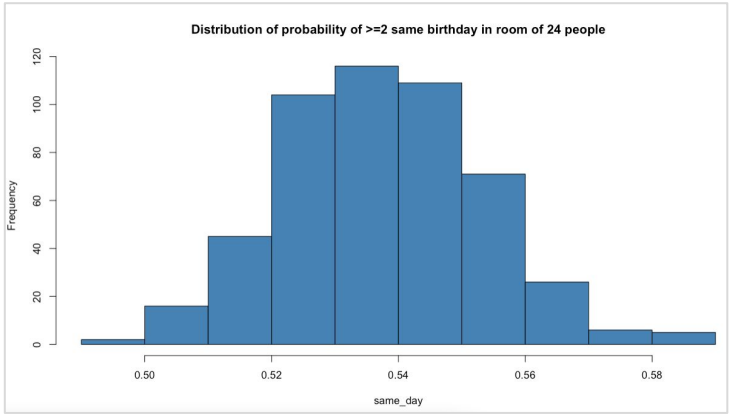
The Birthday Problem

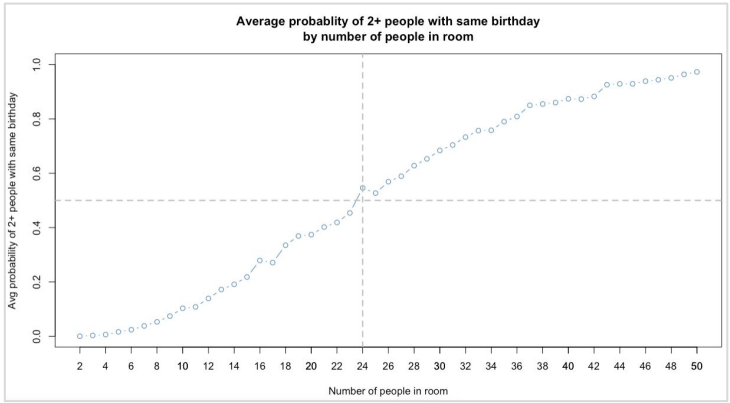
Given a room with 24 randomly selected people, what is the probability that at least two have the same birthday?

How does the probability of two or more people having the same birthday change as a function of the number of people in the room?

The Birthday Problem

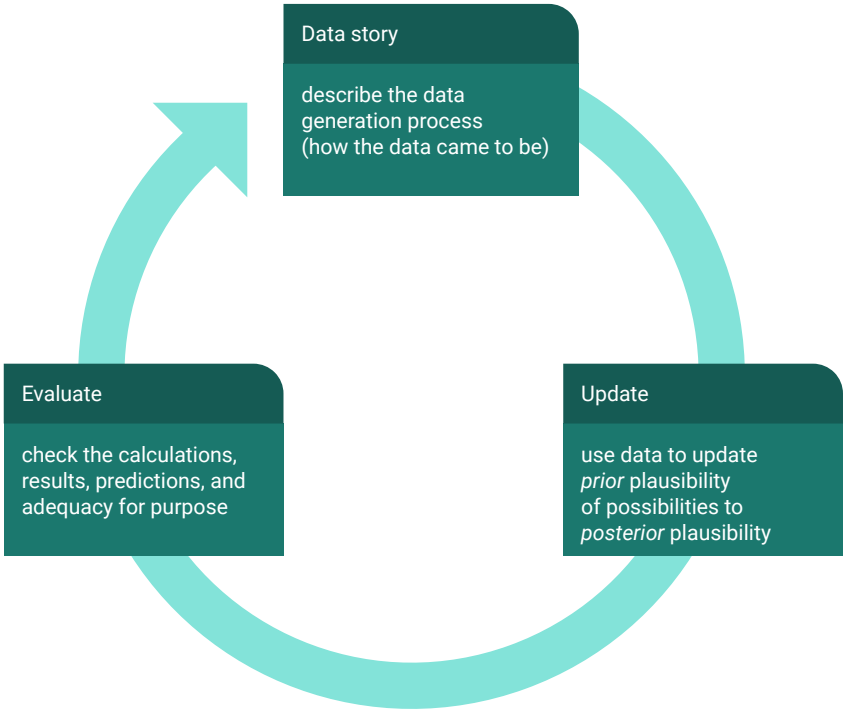
```
> sims <- 1000
> people <- 24
> all_days <- seq(1, 365, 1)
> same_day <- 0
>
> for (i in 1:sims) {
+   room <- sample(x = all_days, size = people, replace = TRUE)
+   if( length( unique(room) ) < people ) same_day <- same_day + 1
+ }
>
> cat("Prob(at least two with same birthday):", same_day / sims, "\n")
Prob(at least two with same birthday): 0.548
```





Introduction to Regression, cont.

How do we design the model? Basic model design loop.



How do we design the model? Basic model design loop.

Data story: describe the data generation process (how the data came to be)

- Descriptive
- Causal

Update: use data to update prior plausibility of possibilities to posterior plausibility

- Prior
- Data
- Posterior

Evaluate: check the calculations, results, predictions, and adequacy for purpose

- Supervise
- Critique

Modify and repeat (potentially)

Linear regression

Let’s review the components of our linear model (y = a + bx + error):

$y_i \sim \text{Normal}(\mu_i, \sigma)$ [likelihood]

$\mu_i = a + b(x_i - \text{mean}(x))$ [linear model]

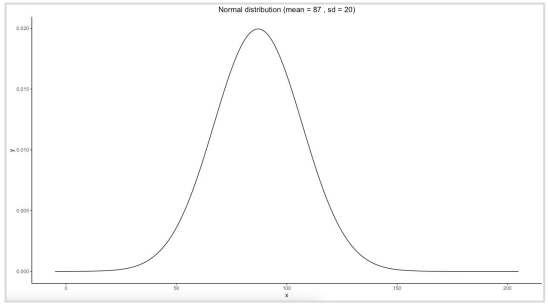
$a \sim \text{Normal}(87, 20)$ [a prior]

$b \sim \text{Normal}(0, 10)$ [b prior]

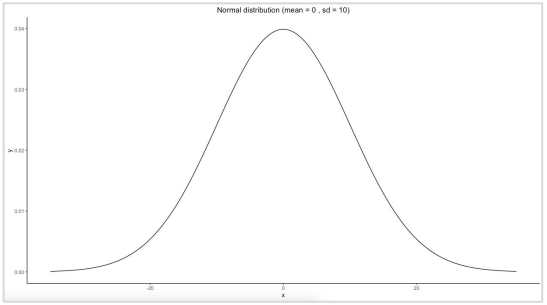
$\sigma \sim \text{Exponential}(1)$ [σ prior]

Prior distributions for the parameters

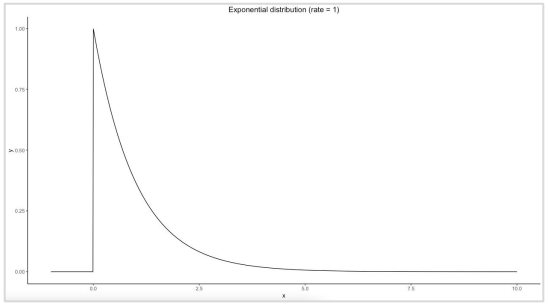
a



b

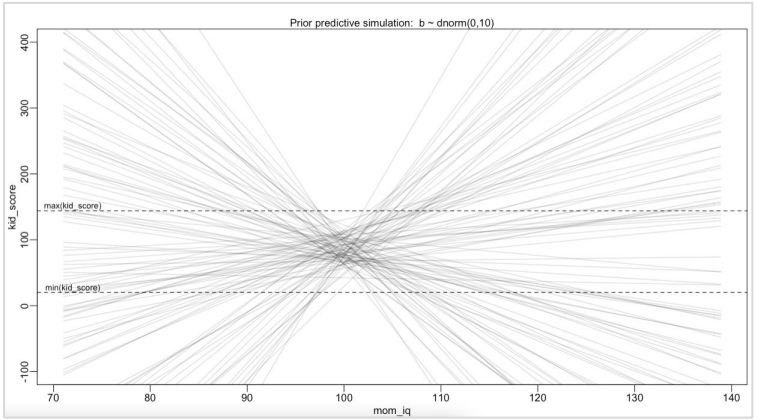


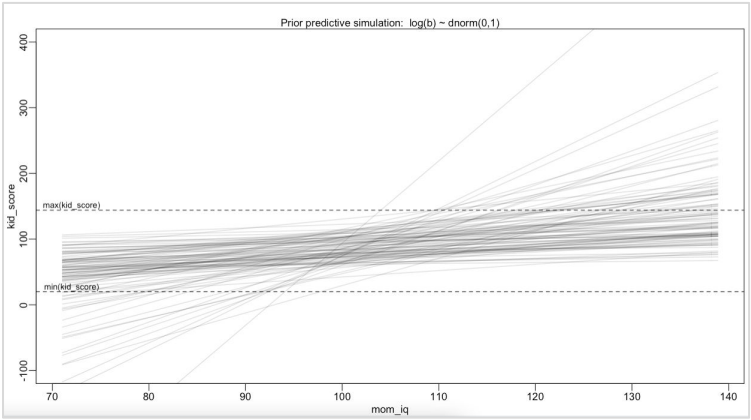
σ



Linear regression: Prior predictive simulation

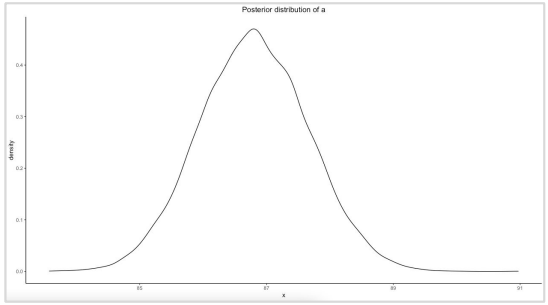
A Gaussian prior centered on zero places as much probability below zero as above zero, and when b = 0, mom_iq has no relationship to kid_score. This prior is too flexible, but in this case the data will overwhelm it. Let's do better:



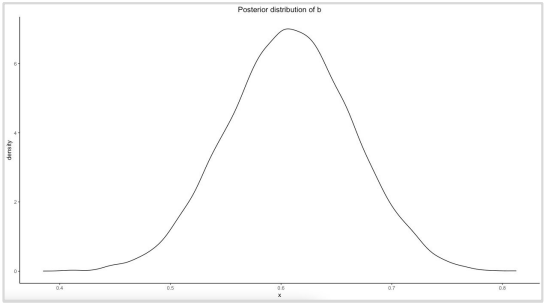


Posterior distributions for the parameters

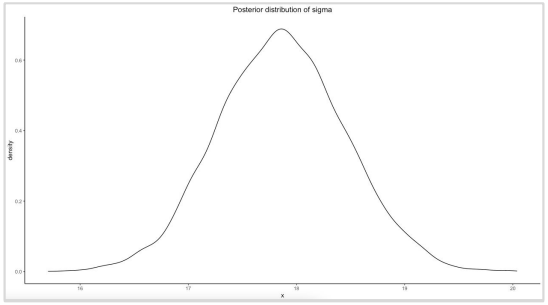
a



b



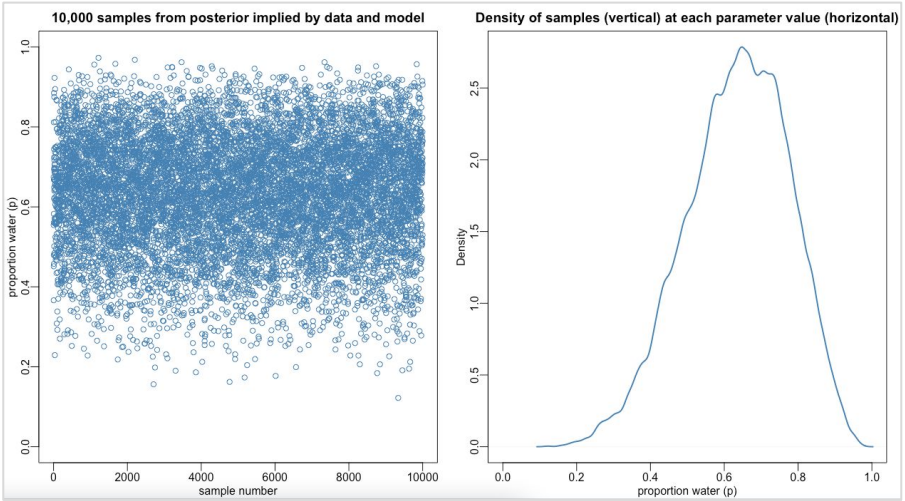
σ



Sampling from the posterior distribution

Prior predictive simulation

Posterior predictive simulation



Sampling to summarize

Once your model produces a posterior distribution, the model's work is done. But your work has just begun. It is necessary to summarize and interpret the posterior distribution. Exactly how it is summarized depends upon your purpose. But common questions include:

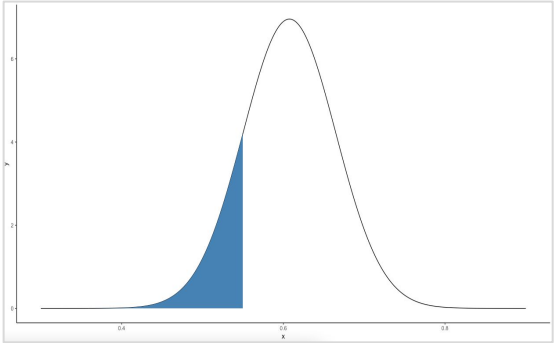
- How much posterior probability lies below some parameter value?
- How much posterior probability lies between two parameter values?
- Which parameter value marks the lower 5% of the posterior probability?
- Which range of parameter values contains 90% of the posterior probability?
- Which parameter value has highest posterior probability?

These simple questions can be usefully divided into questions about (1) intervals of ***defined boundaries***, (2) questions about intervals of ***defined probability mass***, and (3) questions about ***point estimates***. Let's see how to approach these questions using samples from the posterior

Intervals of defined boundaries

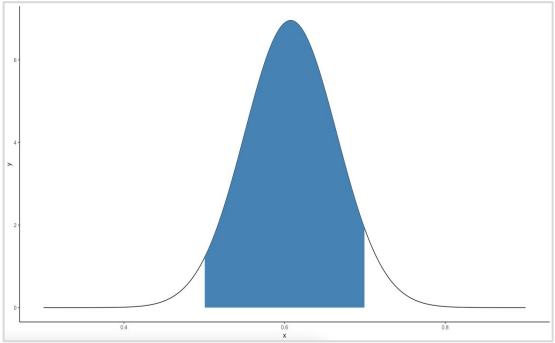
Suppose I ask you for the posterior probability that the slope coefficient is less than 0.55. To calculate this probability, add up all of the samples below 0.55 and divide the resulting count by the total number of samples. In other words, find the frequency of parameter values below 0.55:

```
> sum( samples$b < 0.55 ) / length(samples$b)
[1] 0.1578
```



Using the same approach, we can ask how much posterior probability lies between 0.5 and 0.7:

```
> sum( samples$b > 0.5 & samples$b < 0.7 ) / length(samples$b)
[1] 0.9183
```



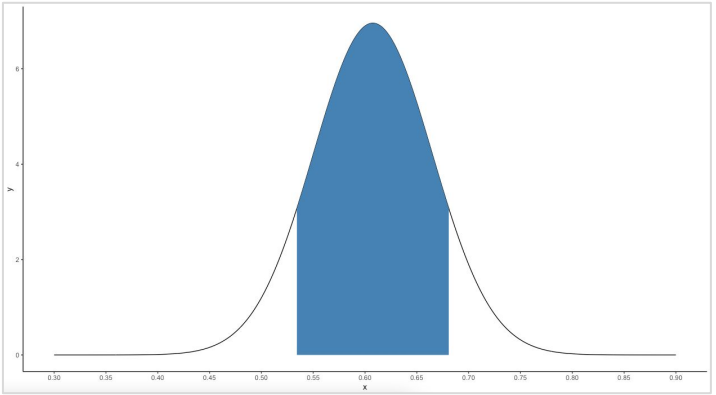
So about 92% of the posterior probability for b lies between 0.5 and 0.7.

Intervals of defined mass

It's common to see researchers report intervals of defined mass, usually known as confidence intervals. An interval of posterior probability, such as the ones we are working with, may instead be called a *credible interval* or *compatibility interval*. These posterior intervals report two parameter values that contain between them a specified amount of posterior probability, a probability mass.

Suppose we want to know the boundaries of the middle 80% posterior probability, which lies between the 10 percentile and the 90th percentile:

```
> quantile( samples$b , c( 0.1 , 0.9 ) )
      10%       90%
0.5348409 0.6816332
```



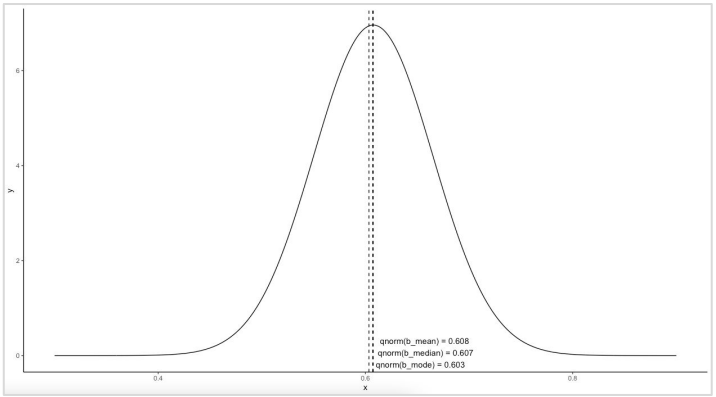
Intervals of this sort, which assign equal probability mass to each tail, are very common. We'll call them *percentile intervals* (PI). They do a good job of communicating the shape of a distribution, as long as the distribution isn't too asymmetrical

Point estimates

Another common task is to produce point estimates. Given the entire posterior distribution, what value should you report? This question isn't easy to answer.

One principled way to go beyond using the entire posterior as the estimate is to choose a *loss function*. Two common ones are *absolute loss*, which leads to the median as the point estimate, and *quadratic loss*, which leads to the mean as the point estimate

```
> round( chainmode( samples$b ) , 3)
[1] 0.603
>
> round( median( samples$b ) , 3)
[1] 0.607
> round( mad( samples$b) , 3)
[1] 0.058
>
> round( mean( samples$b ) , 3)
[1] 0.608
> round( sd( samples$b) , 3)
[1] 0.057
```



However, keep in mind that the Bayesian parameter estimate is the entire posterior distribution, so the important point to note is that you don't have to choose a point estimate. Doing so discards information.

Appendix

Resources

Regression and Other Stories

Statistical Rethinking

Statistical rethinking with brms, ggplot2, and the tidyverse: Second edition

Bayes Rules!

Doing Bayesian Data Analysis, Second edition

Doing Bayesian Data Analysis in brms and the tidyverse

rstanarm vignettes

bayesplot vignettes