Continuous RV, Normal (Gaussian) Distribution

STA 032: Gateway to data science Lecture 17

Jingwei Xiong

May 10, 2023

Recall: Continuous random variables

- We saw continuous random variables earlier and how they differ from discrete random variables

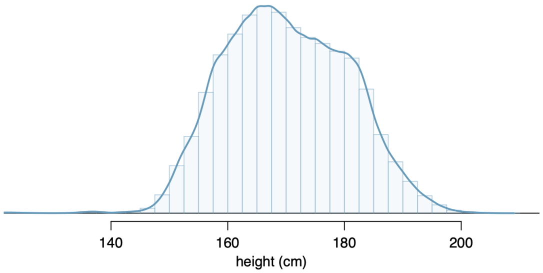


Figure 3.26: The continuous probability distribution of heights for US adults.

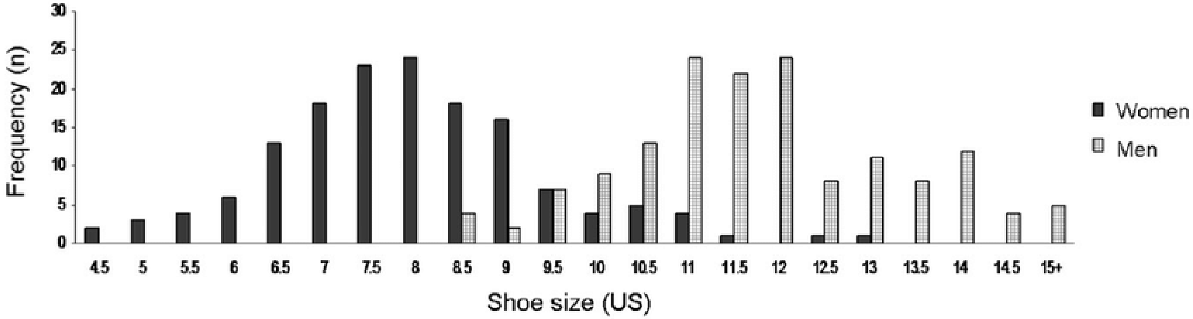
- A probability distribution for a discrete random variable is called a **probability mass function**; for a continuous random variable it is a **probability density functions**
- For a continuous random variable, the probability that the random variable takes on any exact value is zero. Instead, we think about probabilities in ranges.
- $P(a \leq X \leq b)$ is the area under the density function between a and b.

Normal Distribution

- The **normal distribution** is an example of a continuous distribution
- It is a very important distribution and one of the primary inferential tools in statistics
- Many natural phenomenon approximate the normal distribution, such as weight, height, blood pressure, annual rainfall
- The normal distribution is commonly called the *Gaussian distribution* after Carl Friedrich Gauss, who wrote down the equations governing it in the early 1800's.
- It is also sometimes referred to as a *bell curve*, although there are other distributions that are symmetric and shaped like a bell

Illustration: Shoe sizes

- Thinking about shoe sizes is a nice formulation for the Gaussian distribution
- Mickle et al (2010 *Footwear Science*) showed the following bimodal distribution of shoe sizes in the US.



Note that standard shoe sizes are discrete.

Illustration: Shoe sizes

- Let X represent shoe size for wearers of men's shoes
- Here is a (hypothetical) probability distribution of shoe sizes of wearers of men's shoes.

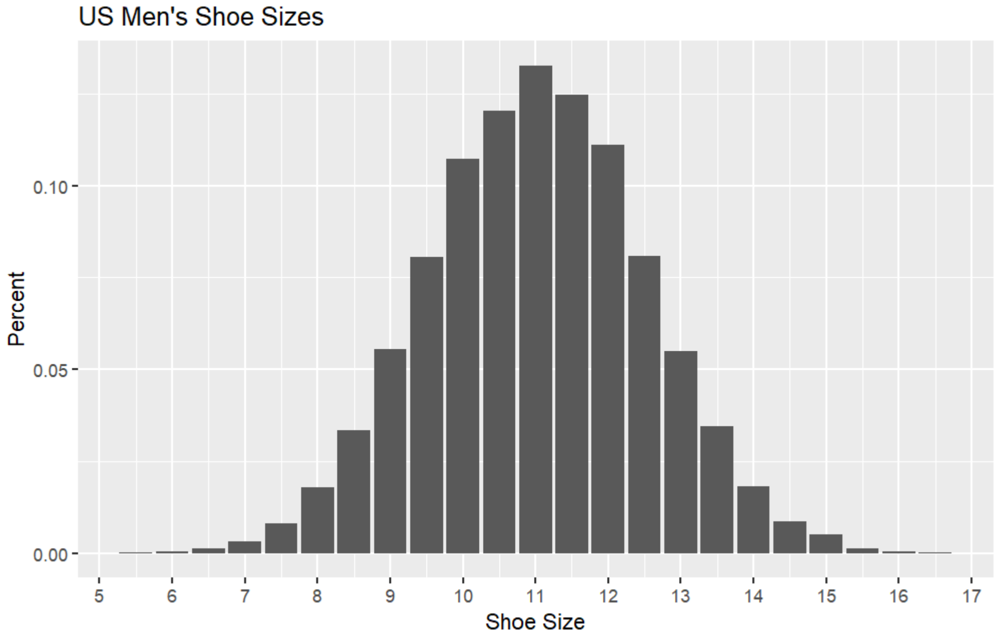
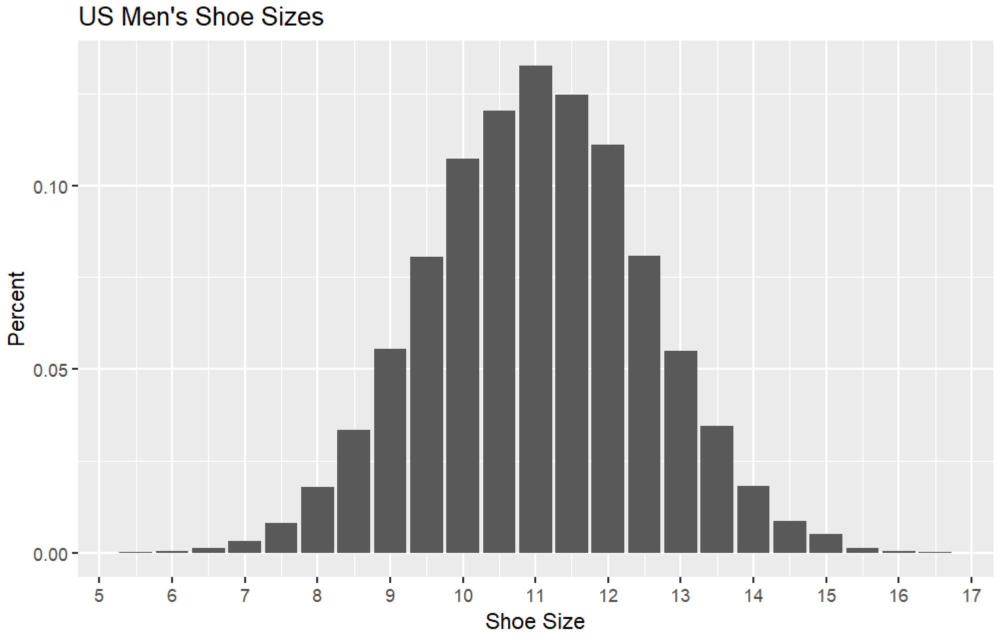


Illustration: Shoe sizes

If we want to know the probability that a customer coming into a store wants a men's shoe size smaller than 9, we just add up the heights of the bars for shoe sizes 8.5 and smaller. We can do this for shoe sizes in any range and tabulate the full discrete distribution of shoe sizes.



Smaller Shoes

```
 size probability
  5.5      0.0001
  6.0      0.0006
  6.5      0.0012
  7.0      0.0032
  7.5      0.0081
  8.0      0.0180
  8.5      0.0334
  9.0      0.0556
  9.5      0.0805
 10.0      0.1072
 10.5      0.1202
```

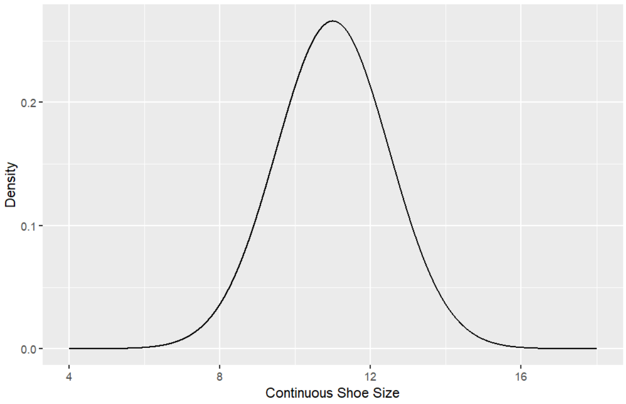
```
 size probability
 11.0      0.1326
 11.5      0.1247
 12.0      0.1109
 12.5      0.0807
 13.0      0.0550
 13.5      0.0345
 14.0      0.0182
 14.5      0.0086
 15.0      0.0050
 15.5      0.0012
 16.0      0.0004
```

The probability of a random men's shoe wearer having a shoe size less than 9 in this population is 0.0646.

What is the probability of shoe size 10-11.5?

Moving to Continuous Distributions

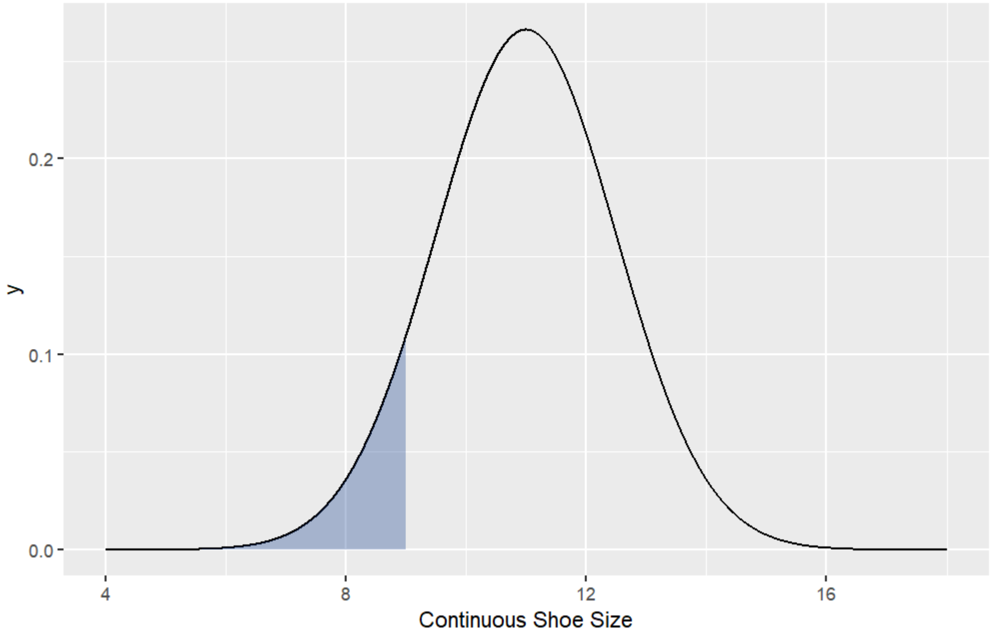
- Now suppose we could get *really* well-fitting shoes, using quarter sizes (9, 9.25, 9.5, 9.75, …) or even tenth sizes (9, 9.1, 9.2, …), or shoes specifically made to fit your feet perfectly.
- As the number of sizes increases, the bar width becomes more narrow, and the graph approaches a smooth curve.
- We will use these smooth curves to describe the probability distributions of continuous random variables (e.g. a shoe size could be 9.50032)



This is a *probability density function*.

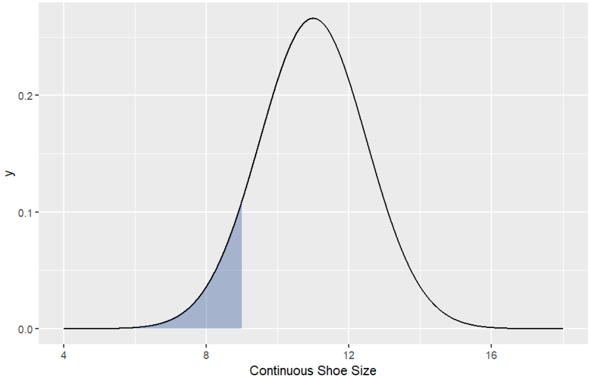
Moving to Continuous Distributions

- The probability density function can be used to get the probability of any range of continuous shoe sizes we would like to investigate.



For example, we can calculate the probability a continuous shoe size is less than 9 (the shaded area).

Moving to Continuous Distributions



- How do we find this area of interest?
- Calculus!

$$P(a \le X \le b) = \text{area between a and b below the curve} = \int_a^b f(x)dx$$

where $f(x)$ represents the density curve

  - In this course, we will use R

Normal Distribution

- The normal distribution is a **symmetric**, **bell-shaped** distribution
- It is characterized by the mean, μ, and the standard deviation, σ (or variance, σ^2)
- For the normal distribution, the **density function** is given by

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

- The notation $N(\mu, \sigma^2)$ is often used.
- The normal distribution with mean 0 and standard deviation 1 is called the **standard normal distribution**. It is commonly denoted $Z \sim N(0, 1)$.

Probability density function for Normal Distribution

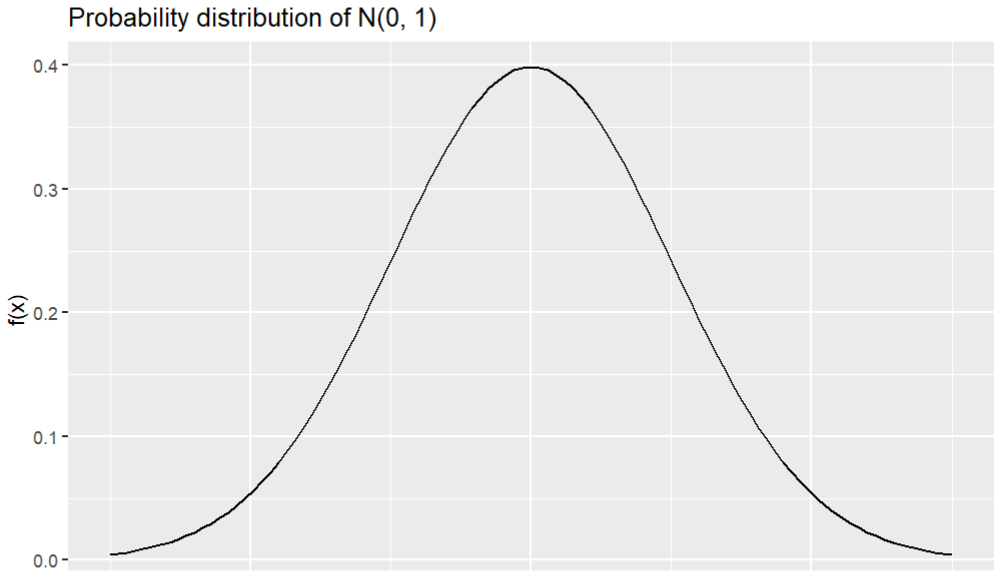
- Like `dbinom()` and `dpois()`, `dnorm()` in R gives us the probability distribution
- Here instead of $P(X = x)$, it is the **value of the probability density function**, $f(x)$ on the previous slide, at values that we input
- `dnorm()` has arguments `x`, `mean` and `sd`, where `mean` and `sd` are the mean and standard deviation of the normal distribution that we want
- **Remember that** $P(X = x) = 0$ **for a continuous random variable**; the value that `dnorm()` gives us is not a probability but the height of the density function

Probability density function for Normal Distribution

```r
dnorm(x = -3:3, mean = 0, sd = 1)
```

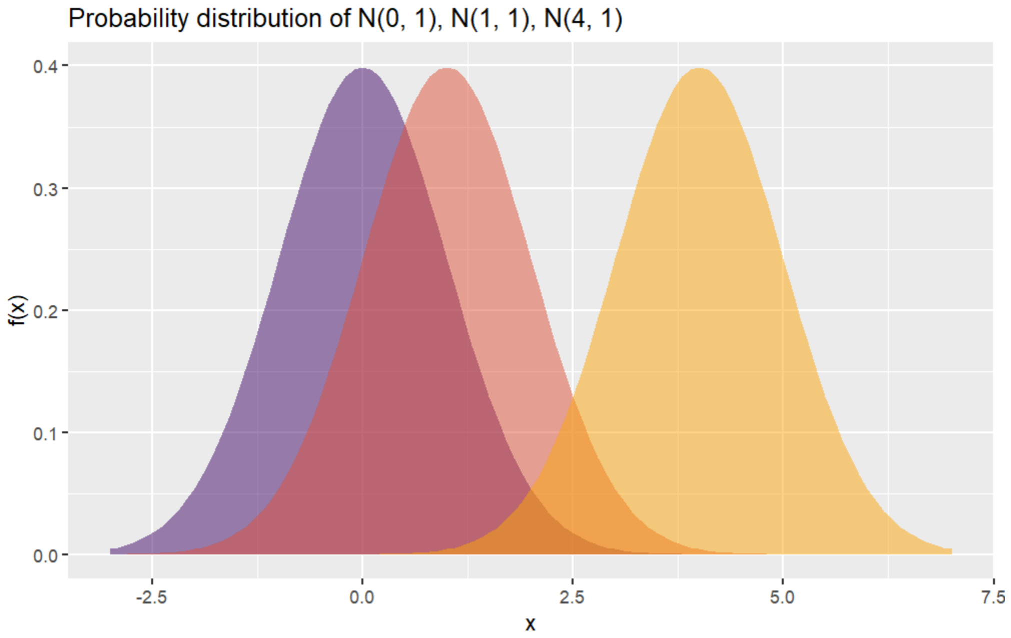
```
[1] 0.004431848 0.053990967 0.241970725 0.398942280 0.241970725 0.053990967
[7] 0.004431848
```

```r
data.frame(x = c(-3, 3)) %>%
  ggplot(aes(x)) +
  stat_function(fun = dnorm, args = list(mean = 0, sd = 1)) +
  labs(title = "Probability distribution of N(0, 1)",
       y = "f(x)")
```



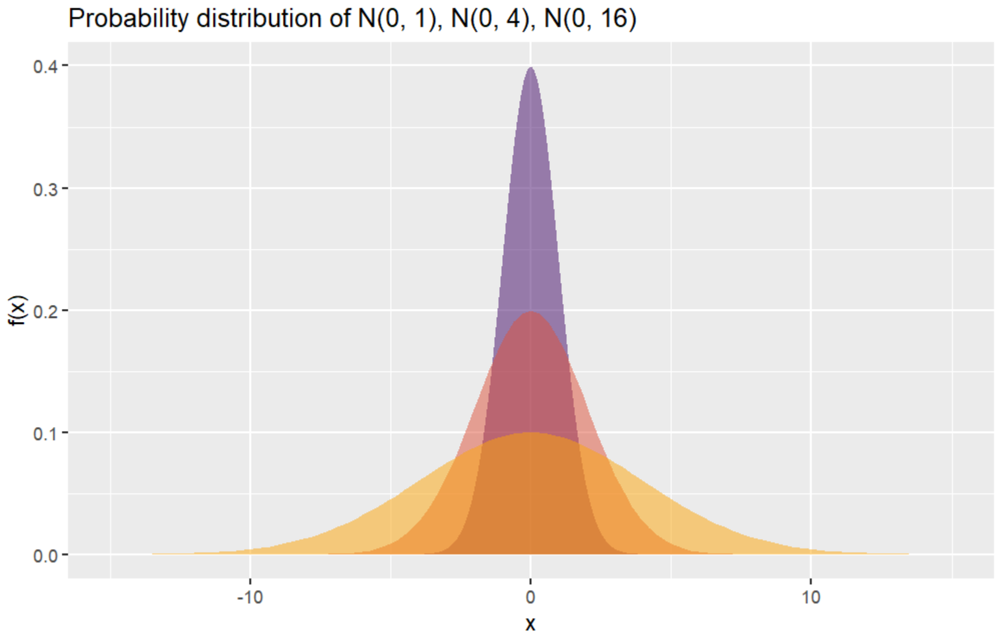
Normal Distribution varying mean

- Which of the three distributions have means 0, 1, and 4?



Normal Distribution varying standard deviation

- Which has standard deviation 1, 2, and 4?



Calculating probabilities for the normal distribution

- We saw `pbinom()`, which gave us $P(X \leq x)$ for the binomial distribution
- Similarly, `pnorm()` gives us $P(X \leq x)$ for the normal distribution. The arguments are
  - `q`, the vector of quantiles (x in $P(X \leq x)$); note that you can input multiple values at once, hence "vector"
  - `mean`, the mean μ (default value 0)
  - `sd`, the standard deviation σ (default value 1)

```
pnorm(0)
```

```
[1] 0.5
```

Calculating probabilities for our shoes example

Going back to our shoe size example, assume that men's shoe sizes follow a normal distribution with mean 11 and standard deviation 1.5, i.e., $N(\mu = 11, \sigma^2 = 1.5^2)$

What is the probability of shoe sizes less than 9?

```
pnorm(9, mean = 11, sd = 1.5)
```

```
[1] 0.09121122
```

What is the probability of shoe sizes greater than 9?

```
1 - pnorm(9, mean = 11, sd = 1.5)
```

```
[1] 0.9087888
```

Calculating probabilities for our shoes example

What is the probability of shoe sizes less than 13?

```r
pnorm(13, mean = 11, sd = 1.5)
```

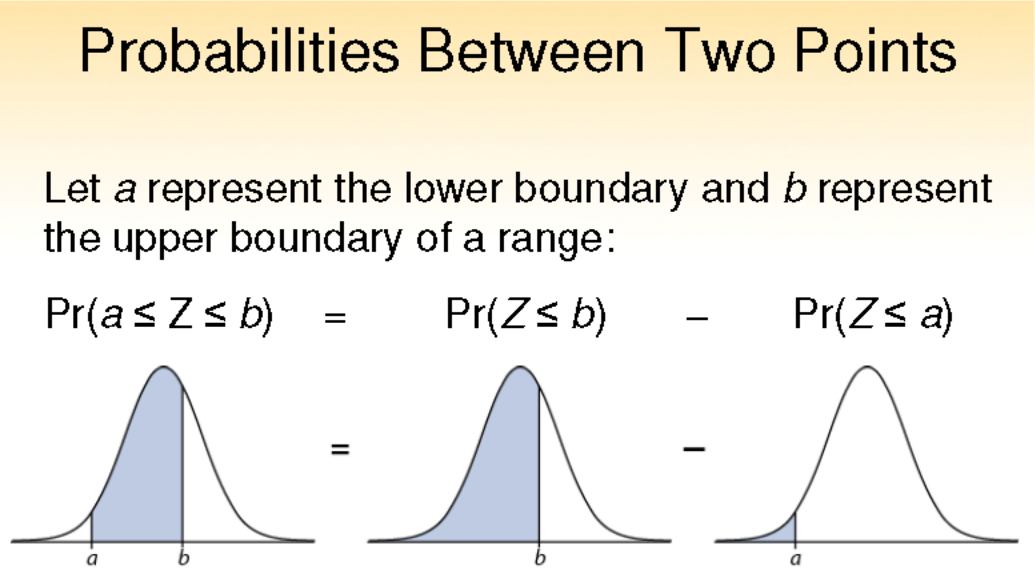
```
[1] 0.9087888
```

What is the probability of shoe size 10-11.5?

```r
pnorm(11.5, mean = 11, sd = 1.5) - pnorm(10, mean = 11, sd = 1.5)
```

```
[1] 0.3780661
```

Probabilities between two values



Probabilities Between Two Points

Let *a* represent the lower boundary and *b* represent the upper boundary of a range:

$\Pr(a \leq Z \leq b) \quad = \quad \Pr(Z \leq b) \quad - \quad \Pr(Z \leq a)$

To get the probability that a random wearer of men's shoes would wear a size between 10 and 11.5, we take `pnorm(11.5, mean = 11, sd = 1.5) - pnorm(10, mean = 11, sd = 1.5)` to get the value 0.3780661.

Sampling from Normal distribution in R

- Just like with the Bernoulli, binomial and Poisson distributions, we can simulate random draws from the normal distribution using the `rnorm()` function

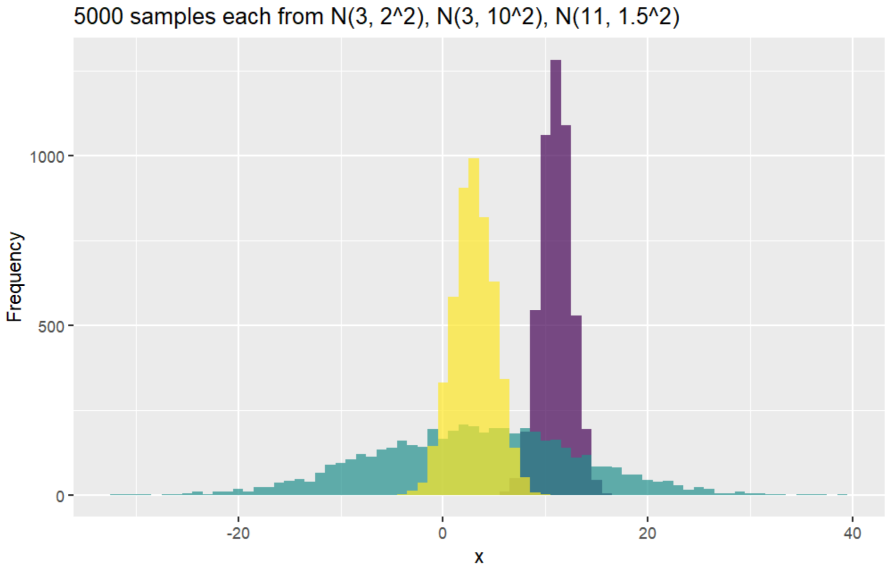
- `rnorm()` has the arguments `n, mean, sd`, where `n` is the number of draws from the distribution, `mean` is the mean and `sd` is the standard deviation.

```
set.seed(0) # so results are reproducible
normalDraws <- rnorm(n = 100, mean = 0, sd = 1)
head(normalDraws, 20)
```

```
 [1]  1.262954285 -0.326233361  1.329799263  1.272429321  0.414641434
 [6] -1.539950042 -0.928567035 -0.294720447 -0.005767173  2.404653389
[11]  0.763593461 -0.799009249 -1.147657009 -0.289461574 -0.299215118
[16] -0.411510833  0.252223448 -0.891921127  0.435683299 -1.237538422
```

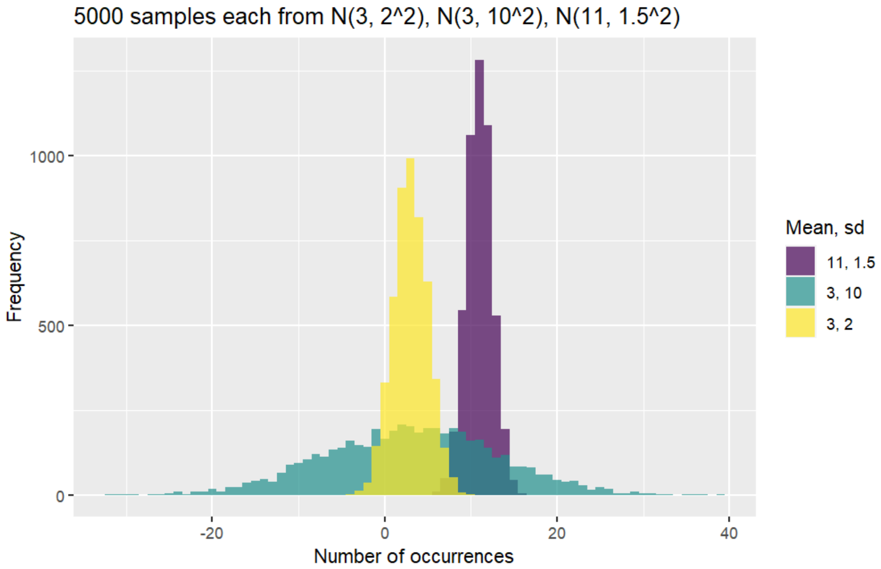
Frequency distribution varying mean and sd

```r
set.seed(0) # so results are reproducible
normal1 <- rnorm(n = 5000, mean = 3, sd = 2)
normal2 <- rnorm(n = 5000, mean = 3, sd = 10)
normal3 <- rnorm(n = 5000, mean = 11, sd = 1.5) # shoe size distribution
```



Frequency distribution varying mean and sd

```
set.seed(0) # so results are reproducible
normal1 <- rnorm(n = 5000, mean = 3, sd = 2)
normal2 <- rnorm(n = 5000, mean = 3, sd = 10)
normal3 <- rnorm(n = 5000, mean = 11, sd = 1.5)
```



More about the standard normal distribution

- Recall: standard normal distribution $Z \sim N(0, 1)$
- A normally distributed random variable can be expressed as a standard normal by **subtracting the mean and dividing by the standard deviation**; this process is called **standardization**
- $Y \sim N(\mu, \sigma^2)$
- $Z = \frac{Y-\mu}{\sigma}$
- $E\left(\frac{Y-\mu}{\sigma}\right) = \frac{1}{\sigma}[E(Y) - \mu] = 0$
- $Var\left(\frac{Y-\mu}{\sigma}\right) = \frac{1}{\sigma^2}[Var(Y)] = \frac{1}{\sigma^2}[\sigma^2] = 1$
- What we are essentially doing is **moving the location** (mean moves to 0) and **changing the scale** (standard deviation becomes 1)

More about the standard normal distribution

- Earlier, we were interested in the probability of shoe sizes smaller than 13, and we calculated it using

```
pnorm(13, mean = 11, sd = 1.5)
```

```
[1] 0.9087888
```

- Let Y be the random variable denoting men's shoe sizes. Then $Y \sim N(11, 1.5^2)$.

$$
\begin{aligned}
P(Y \leq 13) &= P\left(\frac{Y - \mu_y}{\sigma_y} \leq \frac{13 - \mu_y}{\sigma_y}\right) \\
&= P\left(Z \leq \frac{13 - 11}{1.5}\right) \\
&= P(Z \leq \frac{2}{1.5})
\end{aligned}
$$

```
pnorm(2/1.5, mean = 0, sd = 1)
```

```
[1] 0.9087888
```

z-score

$$
\begin{aligned}
P(Y \leq 13) &= P\left(\frac{Y - \mu_y}{\sigma_y} \leq \frac{13 - \mu_y}{\sigma_y}\right) \\
&= P\left(Z \leq \frac{13 - 11}{1.5}\right) \\
&= P(Z \leq \frac{2}{1.5})
\end{aligned}
$$

- The standardized value on the right-hand side, $\frac{13-11}{1.5}$, is known more generally as a z-score, where $z = \frac{x-\mu}{\sigma} = \frac{\text{value - mean}}{\text{standard deviation}}$
- The z-score is the **number of standard deviations above (positive z-scores) or below the mean (negative z-scores)**. To see this:
  - $x - \mu$ is the number relative to the mean, e.g., shoe size 13 is 2 above the mean
  - Dividing the above by σ gives us the number of standard deviations above the mean, e.g., with our shoe size distribution having a standard deviation of 1.5, shoe size 13 is $\frac{2}{1.5} = 1.33$ standard deviations above the mean

z-score

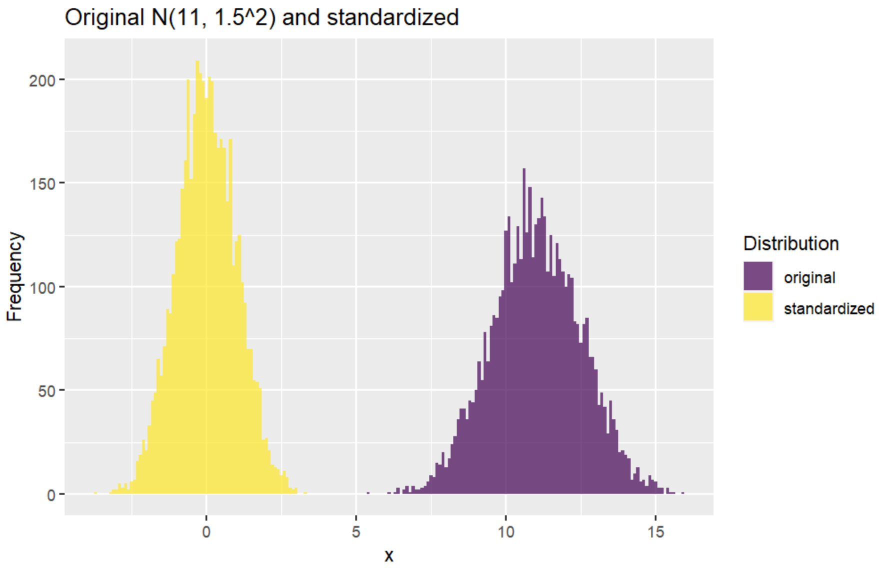
$$
\begin{aligned}
P(Y \leq 13) &= P\left(\frac{Y-\mu_y}{\sigma_y} \leq \frac{13-\mu_y}{\sigma_y}\right) \\
&= P\left(Z \leq \frac{13-11}{1.5}\right) \\
&= P(Z \leq \frac{2}{1.5})
\end{aligned}
$$

- The *relative* positions of values in the original and standardized distributions stay the same, i.e., $P(Y \leq 13) = P(Z \leq \frac{2}{1.5})$
  - Our size 13 (or value $\frac{2}{1.5}$ standard deviations above the mean) remains the same relative to the rest of the distribution.

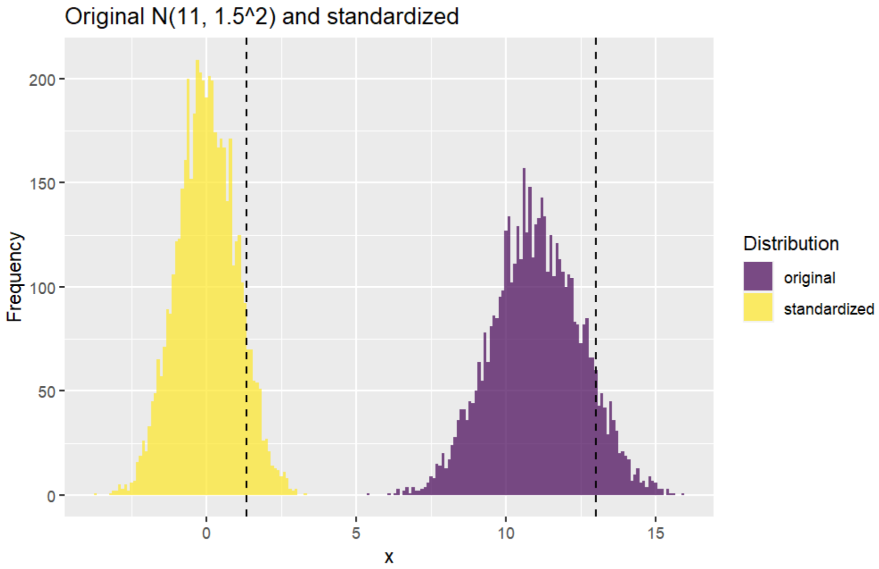
Standardizing in R

Consider the samples we drew earlier from $N \sim (11, 1.5^2)$

```
set.seed(0) # so results are reproducible
normal3 <- rnorm(n = 5000, mean = 11, sd = 1.5) # shoe size distribution
standardizedNormal3 <- (normal3 - 11)/1.5
```



Standardizing in R



```
sum(normal3 <= 13)/length(normal3)
```

```
[1] 0.9072
```

```
sum(standardizedNormal3 <= 2/1.5)/length(standardizedNormal3)
```

```
[1] 0.9072
```

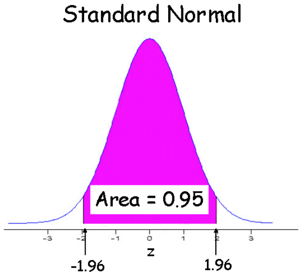
More about the standard normal distribution

- We saw earlier that $P(Z \leq 0) = .5$. This is because the standard normal distribution is symmetric with mean 0.

```
pnorm(0) # default value of mean = 0 and sd = 1
```

```
[1] 0.5
```

- Tail probabilities of the standard normal distribution
  - The symmetry of the normal distribution allows us to calculate the probability of values falling in the tails
  - For any z-score, $P(Z \leq -z) = P(Z \geq z)$



Quantiles for the normal distribution

- Quantiles are cut points dividing the range of a probability distribution into continuous intervals

- Recall: quartiles (four groups) and percentiles (100 groups)

- $P(X \leq q) = p$, where q is the quantile (think of value on the horizontal axis), e.g., $P(Z \leq 0) = .5$

- Recall: `pnorm(q, mean, sd)` for $P(X \leq x)$, or $P(Z \leq z)$ for standard normal. `pnorm()` returns the probability, `p`

```
pnorm(q = 0, mean = 0, sd = 1)
```

```
[1] 0.5
```

- `qnorm(p, mean, sd)` for the quantile, e.g., $P(X \leq \ ?) = p$. `qnorm()` returns the quantile, `q`

```
qnorm(p = .5, mean = 0, sd = 1)
```

```
[1] 0
```

Important reference points for the normal distribution

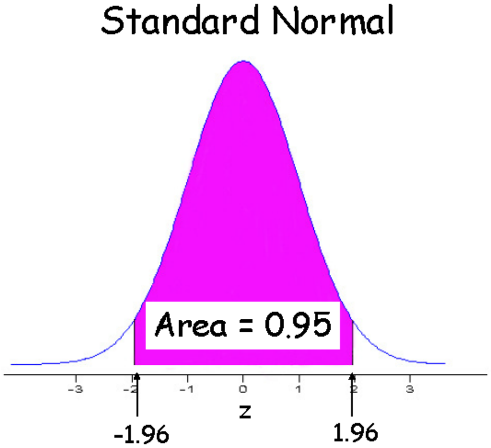
- For the standard normal, z-scores (quantiles in R) corresponding to particular probabilities (critical values) are often written as z_p, where p denotes the probability in the **right tail**, e.g., $z_{.5} = 0$
- The z-scores corresponding to probabilities of 0.025 (2.5%) in the left and right tails are important reference points. Specifically, $z_{.025} \approx 1.96$
- In R, `qnorm(.025, lower.tail = FALSE)` returns the z-score corresponding to a probability of .025 in the right tail, $z_{.025}$, i.e., 2.5% probability in the right tail, so we should get 1.96. By symmetry, `qnorm(.975)` will return 1.96.

```
qnorm(.025, lower.tail = FALSE)
```

```
[1] 1.959964
```

```
qnorm(.975)
```

```
[1] 1.959964
```



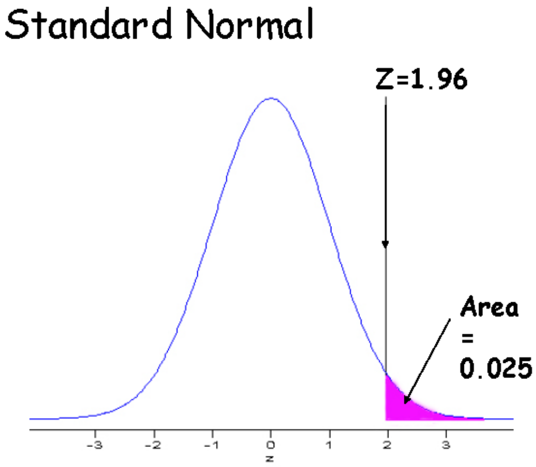
Important reference points for the normal distribution

```
pnorm(1.96)
```

```
[1] 0.9750021
```

```
pnorm(1.96, lower.tail = FALSE)
```

```
[1] 0.0249979
```

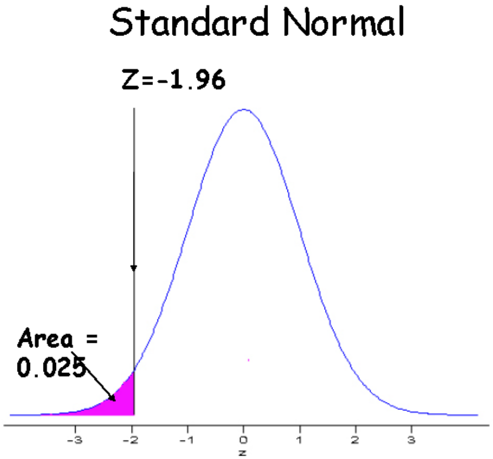


```
pnorm(-1.96)
```

```
[1] 0.0249979
```

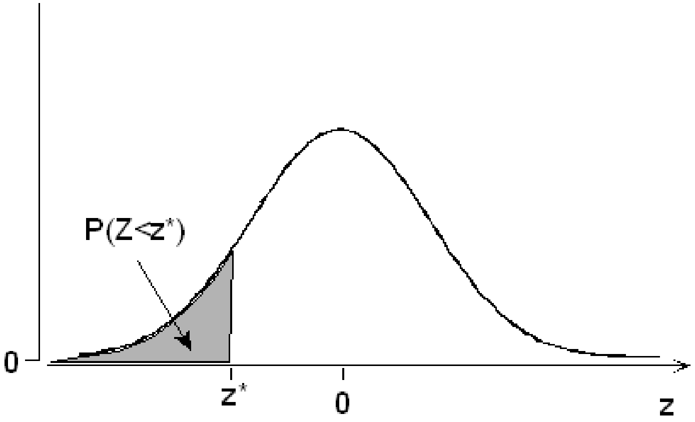
```
pnorm(-1.96, lower.tail = FALSE)
```

```
[1] 0.9750021
```



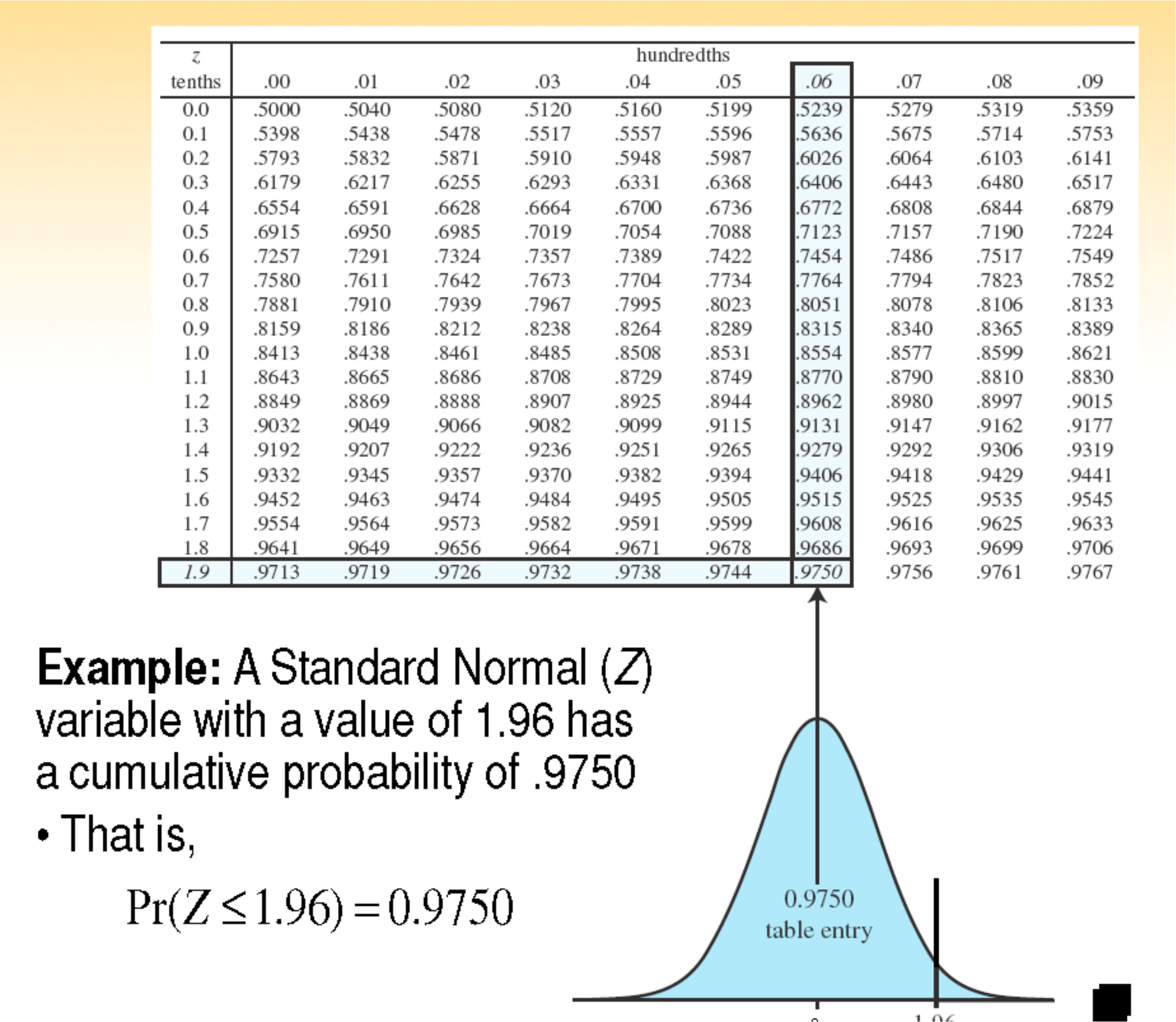
Standard normal table

- A **standard normal table** allows us to calculate values based on the standard normal distribution.
- It tells us how much area is under the normal curve to the *left* of the specified value (lower tail area). Sometimes the table shows the complement of this probability (upper or *right* tail area).



- With modern computing, we don't need to rely on these tables to get the desired probabilities, but you often find them in the back of statistics textbooks.

Standard normal table



z tenths	hundredths .00	 .01	 .02	 .03	 .04	 .05	 .06	 .07	 .08	 .09
0.0	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279	.5319	.5359
0.1	.5398	.5438	.5478	.5517	.5557	.5596	.5636	.5675	.5714	.5753
0.2	.5793	.5832	.5871	.5910	.5948	.5987	.6026	.6064	.6103	.6141
0.3	.6179	.6217	.6255	.6293	.6331	.6368	.6406	.6443	.6480	.6517
0.4	.6554	.6591	.6628	.6664	.6700	.6736	.6772	.6808	.6844	.6879
0.5	.6915	.6950	.6985	.7019	.7054	.7088	.7123	.7157	.7190	.7224
0.6	.7257	.7291	.7324	.7357	.7389	.7422	.7454	.7486	.7517	.7549
0.7	.7580	.7611	.7642	.7673	.7704	.7734	.7764	.7794	.7823	.7852
0.8	.7881	.7910	.7939	.7967	.7995	.8023	.8051	.8078	.8106	.8133
0.9	.8159	.8186	.8212	.8238	.8264	.8289	.8315	.8340	.8365	.8389
1.0	.8413	.8438	.8461	.8485	.8508	.8531	.8554	.8577	.8599	.8621
1.1	.8643	.8665	.8686	.8708	.8729	.8749	.8770	.8790	.8810	.8830
1.2	.8849	.8869	.8888	.8907	.8925	.8944	.8962	.8980	.8997	.9015
1.3	.9032	.9049	.9066	.9082	.9099	.9115	.9131	.9147	.9162	.9177
1.4	.9192	.9207	.9222	.9236	.9251	.9265	.9279	.9292	.9306	.9319
1.5	.9332	.9345	.9357	.9370	.9382	.9394	.9406	.9418	.9429	.9441
1.6	.9452	.9463	.9474	.9484	.9495	.9505	.9515	.9525	.9535	.9545
1.7	.9554	.9564	.9573	.9582	.9591	.9599	.9608	.9616	.9625	.9633
1.8	.9641	.9649	.9656	.9664	.9671	.9678	.9686	.9693	.9699	.9706
1.9	.9713	.9719	.9726	.9732	.9738	.9744	*.9750*	.9756	.9761	.9767

Example: A Standard Normal (*Z*) variable with a value of 1.96 has a cumulative probability of .9750

- That is,

$$\Pr(Z \leq 1.96) = 0.9750$$

Standard normal table

What is the probability of a shoe size bigger than 13 (z-score 1.33)?

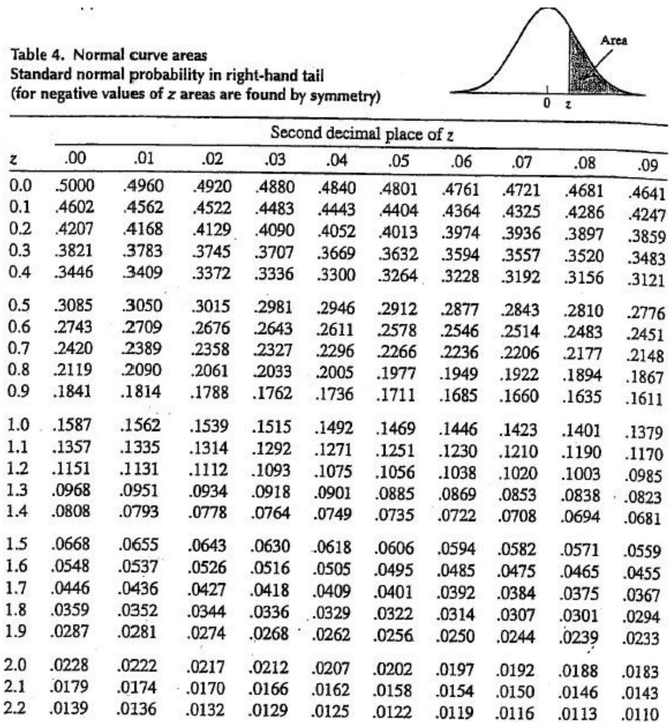


Table 4. Normal curve areas
Standard normal probability in right-hand tail
(for negative values of *z* areas are found by symmetry)

	Second decimal place of z									
z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.5000	.4960	.4920	.4880	.4840	.4801	.4761	.4721	.4681	.4641
0.1	.4602	.4562	.4522	.4483	.4443	.4404	.4364	.4325	.4286	.4247
0.2	.4207	.4168	.4129	.4090	.4052	.4013	.3974	.3936	.3897	.3859
0.3	.3821	.3783	.3745	.3707	.3669	.3632	.3594	.3557	.3520	.3483
0.4	.3446	.3409	.3372	.3336	.3300	.3264	.3228	.3192	.3156	.3121
0.5	.3085	.3050	.3015	.2981	.2946	.2912	.2877	.2843	.2810	.2776
0.6	.2743	.2709	.2676	.2643	.2611	.2578	.2546	.2514	.2483	.2451
0.7	.2420	.2389	.2358	.2327	.2296	.2266	.2236	.2206	.2177	.2148
0.8	.2119	.2090	.2061	.2033	.2005	.1977	.1949	.1922	.1894	.1867
0.9	.1841	.1814	.1788	.1762	.1736	.1711	.1685	.1660	.1635	.1611
1.0	.1587	.1562	.1539	.1515	.1492	.1469	.1446	.1423	.1401	.1379
1.1	.1357	.1335	.1314	.1292	.1271	.1251	.1230	.1210	.1190	.1170
1.2	.1151	.1131	.1112	.1093	.1075	.1056	.1038	.1020	.1003	.0985
1.3	.0968	.0951	.0934	.0918	.0901	.0885	.0869	.0853	.0838	.0823
1.4	.0808	.0793	.0778	.0764	.0749	.0735	.0722	.0708	.0694	.0681
1.5	.0668	.0655	.0643	.0630	.0618	.0606	.0594	.0582	.0571	.0559
1.6	.0548	.0537	.0526	.0516	.0505	.0495	.0485	.0475	.0465	.0455
1.7	.0446	.0436	.0427	.0418	.0409	.0401	.0392	.0384	.0375	.0367
1.8	.0359	.0352	.0344	.0336	.0329	.0322	.0314	.0307	.0301	.0294
1.9	.0287	.0281	.0274	.0268	.0262	.0256	.0250	.0244	.0239	.0233
2.0	.0228	.0222	.0217	.0212	.0207	.0202	.0197	.0192	.0188	.0183
2.1	.0179	.0174	.0170	.0166	.0162	.0158	.0154	.0150	.0146	.0143
2.2	.0139	.0136	.0132	.0129	.0125	.0122	.0119	.0116	.0113	.0110

```r
pnorm(13, mean = 11, sd = 1.5, lower
```

```
[1] 0.09121122
```

```r
pnorm(2/1.5, lower.tail = FALSE)
```

```
[1] 0.09121122
```

```r
1 - pnorm(2/1.5)
```

```
[1] 0.09121122
```

Sum of independent normal random variables

- Important property: Any linear combination of normal random variables is a normal random variable with expectation and variance given by the formulas for expected value and variance of linear combinations (Lecture 15)

- Recall: A linear combination of two random variables, X and Y, is of the form $aX + bY$, where a and b are constants

- Recall:

    - $E(aX + bY) = aE(X) + bE(Y)$
    - For a linear combination of **independent** random variables $Var(aX + bY) = a^2Var(X) + b^2Var(Y)$

- Hence if $X \sim N(\mu_x, \sigma_x^2)$ and $Y \sim N(\mu_y, \sigma_y^2)$ are independent, $W = X + Y \sim N(\mu_x + \mu_y, \sigma_x^2 + \sigma_y^2)$

- Extends to more than two random variables in the linear combination. Note also that b can be negative, e.g., $E(X - Y) = E(X) - E(Y)$ and $Var(X - Y) = Var(X) + Var(Y)$.

Summary: Distributions in R

- For each distribution, R has a family of commands, starting with the letters `d`, `p`, `q` and `r`
  - `d` for density
  - `p` for cumulative density up to input value $P(X \leq x)$. Think of $P(X \leq q) = p$
  - `q` for the quantile, e.g., $P(X \leq \ ?) = p$
  - `r` for a random sample from the distribution

Summary

- Common probability distributions: Normal
  - Theoretical properties: probability density function, parameters, mean and variance, effect of varying parameters
  - R functions:
    - `dnorm()` for densities
    - `pnorm()` for $P(X \leq x)$
    - `rnorm()` for random sample
  - Standard normal distribution