Probability Distributions and Introduction to Statistical Inference

BIO5312 FALL2017

STEPHANIE J. SPIELMAN, PHD

Random variable

Random processes produce numerical outcomes:

- Number of tails in 50 coin flips
- The sum of everyone's heights

Definition: a random variable is a function that maps outcomes of a random process to a numeric value

- X is a function (rule) that assign a number $X(s)$ to each outcome $s \in S$ (where s is an event in sample space S)
- r.v.'s are technically neither random nor variables…
- But, you can think of them roughly numerical outcomes of random processes

Discrete vs continuous RV

Discrete random variables can take on (map to) a finite number of values

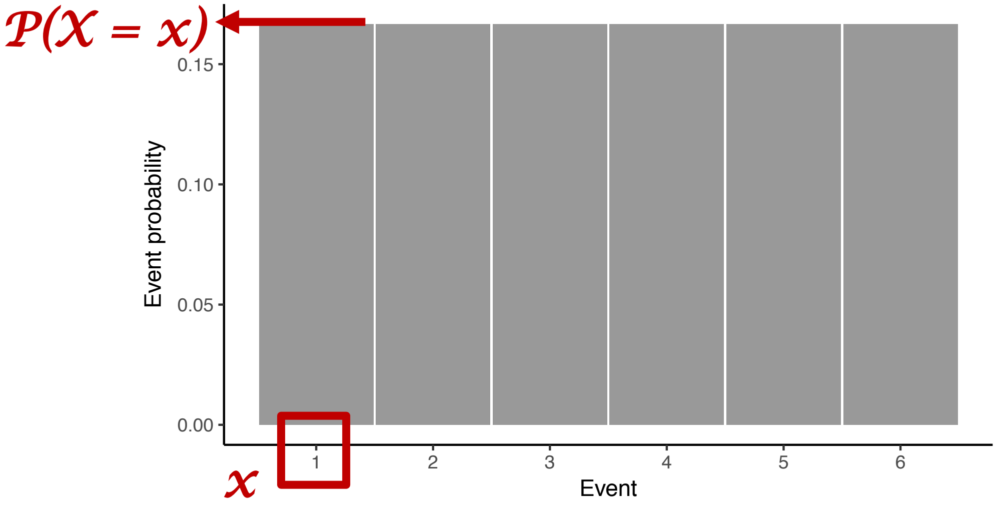
Continuous random variables can take on (map to) innumerable/infinite values

Expressing discrete random variables

Probability mass function (PMF)

- Describes the values taken by a discrete r.v. X and its associated probabilities
- Function that assigns, to any possible value x of a discrete r.v. X, the probability $\mathcal{P}(X = x)$

PMF for rolling a fair die



PMF properties

$0 \ \leq P(X = x) \leq 1$

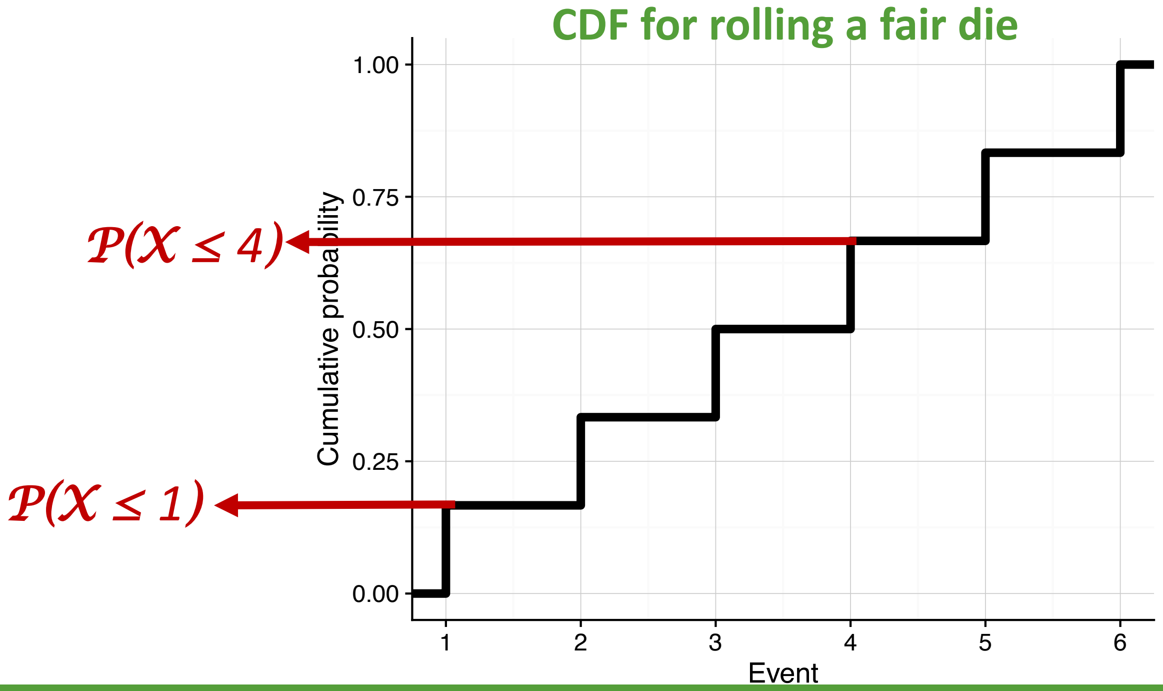
$\sum P(X = x) \ = 1$

PMF is simply a fancier term for a discrete probability distribution

Expressing discrete random variables

Cumulative distribution function (CDF)

- Function defined, for a specific value x of a discrete r.v. $\mathcal{X}$, as $\mathcal{F}(x) = \mathcal{P}(\mathcal{X} \le x)$



CDF properties

$0 \ \leq F(X) \leq 1$

CDF functions are non-decreasing

PMF vs CDF

PMF: What is the probability of event X?

CDF: What is the sum of probabilities for all events ≤ X?

Expectation and spread of random variables

The **expectation** of a r.v. is the probability-weighted average of all possible values (i.e., mean)

- $\mathbb{E}(X) = \mu = \sum_i x_i \, p(x_i)$

The **variance** of a r.v. is defined

- $Var(X) = \sigma^2 = \mathbb{E}[(X - \mu)^2] = \sum_i [x_i^2 p(x_i)] - \mu^2$
- $Var(X) = \mathbb{E}[X^2] - \mathbb{E}[X]^2$

Example: The Binomial distribution

The **binomial distribution** describes the probability of obtaining *k* successes in *n* Bernoulli trials, where the probability of success for each trial is constant at *p*

A **Bernoulli trial** has a binary outcome (success/fail, true/false, yes/no), and P(success) = *p* is the same for all realizations of the trial

The **BInS** conditions

To be binomially distributed, must satisfy the following:

Binary outcomes

Independent trials (outcomes do not influence each other)

n is fixed before the trials begin

Same probability of success, p, for all trials

Is it binomial?

A bag contains 10 balls, 7 red and 3 green.

Situation 1: You draw 5 balls from the bag, noting the ball color each time and then returning it to the bag. **Yes!**

Situation 2: You draw 5 balls from the bag, retaining each drawn ball for safe-keeping so you can play catch at any moment. **No** ☹

Situation 3: You keep drawing balls, with replacement, until you have drawn 4 red balls. **No** ☹

The binomial distribution

The PMF (probability distribution) for a binomially-distributed random variable:

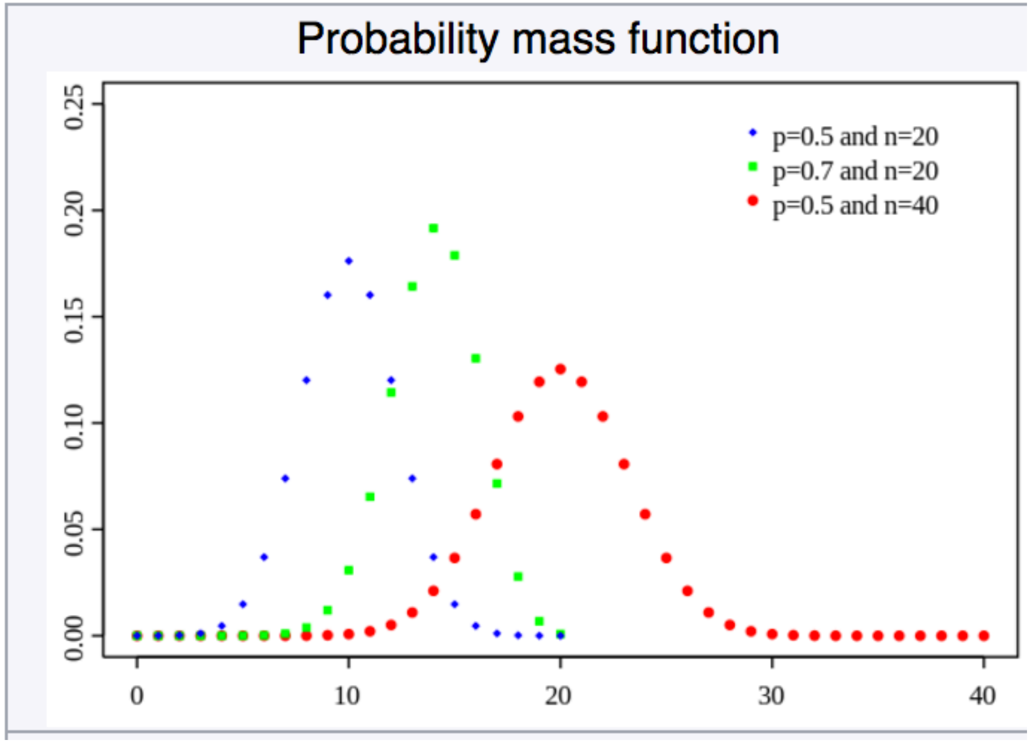
$$P(X = k) = \binom{n}{k} p^k (1-p)^{(n-k)} = \binom{n}{k} p^k q^{(n-k)}$$

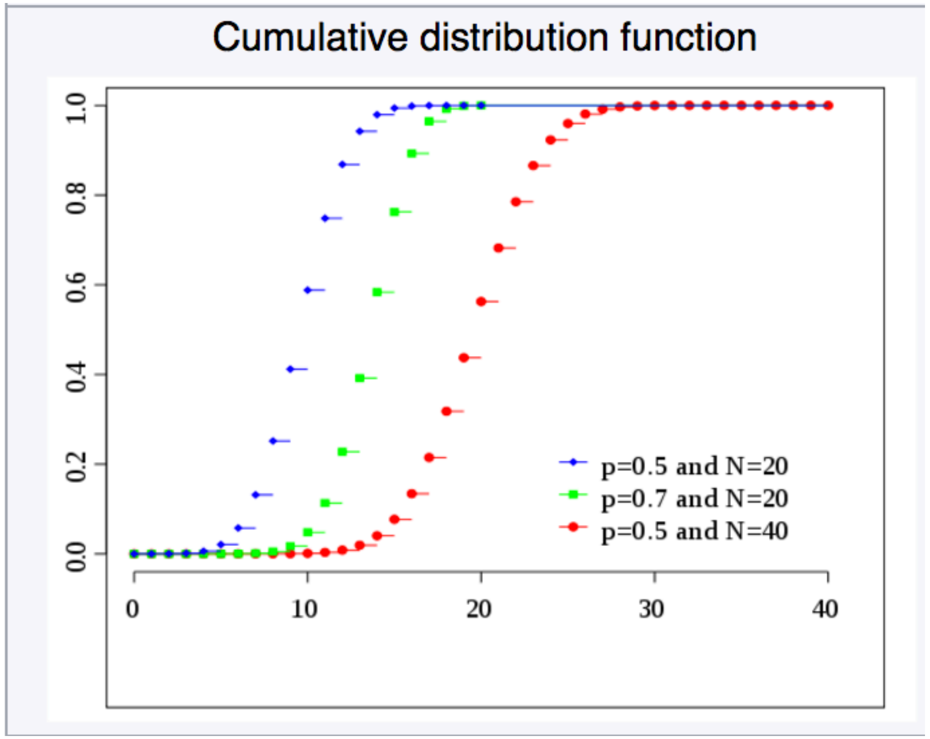
The **binomial coefficient:** $\binom{n}{k} = \frac{n!}{k!(n-k)!}$

- read as "n choose k"

Wikipedia weighs in

binomial





The binomial distribution

The **expectation** for a binomial r.v.

- $\mathbb{E}(X) = \ \mu = \mathrm{np}$

The **variance** for a binomial r.v.

- $Var(X) = \sigma^2 = \mathrm{npq} = \mathrm{np}(1 - \mathrm{p})$

We write binomially distributed r.v.'s as $X \sim B(n, p)$

Example: Playing with a binomial rv

Each child born to a particular set of parents has a 25% probability of having blood type O. Assume the parents had five children.

Here, n = 5 and p = 0.25, meaning we define Type O as "success", and not Type O as "failure". → X~B(5, 0.25)

Tasks:

- Compute expectation and variance
- Visualize PMF
- Visualize CDF
- Make some calculations…

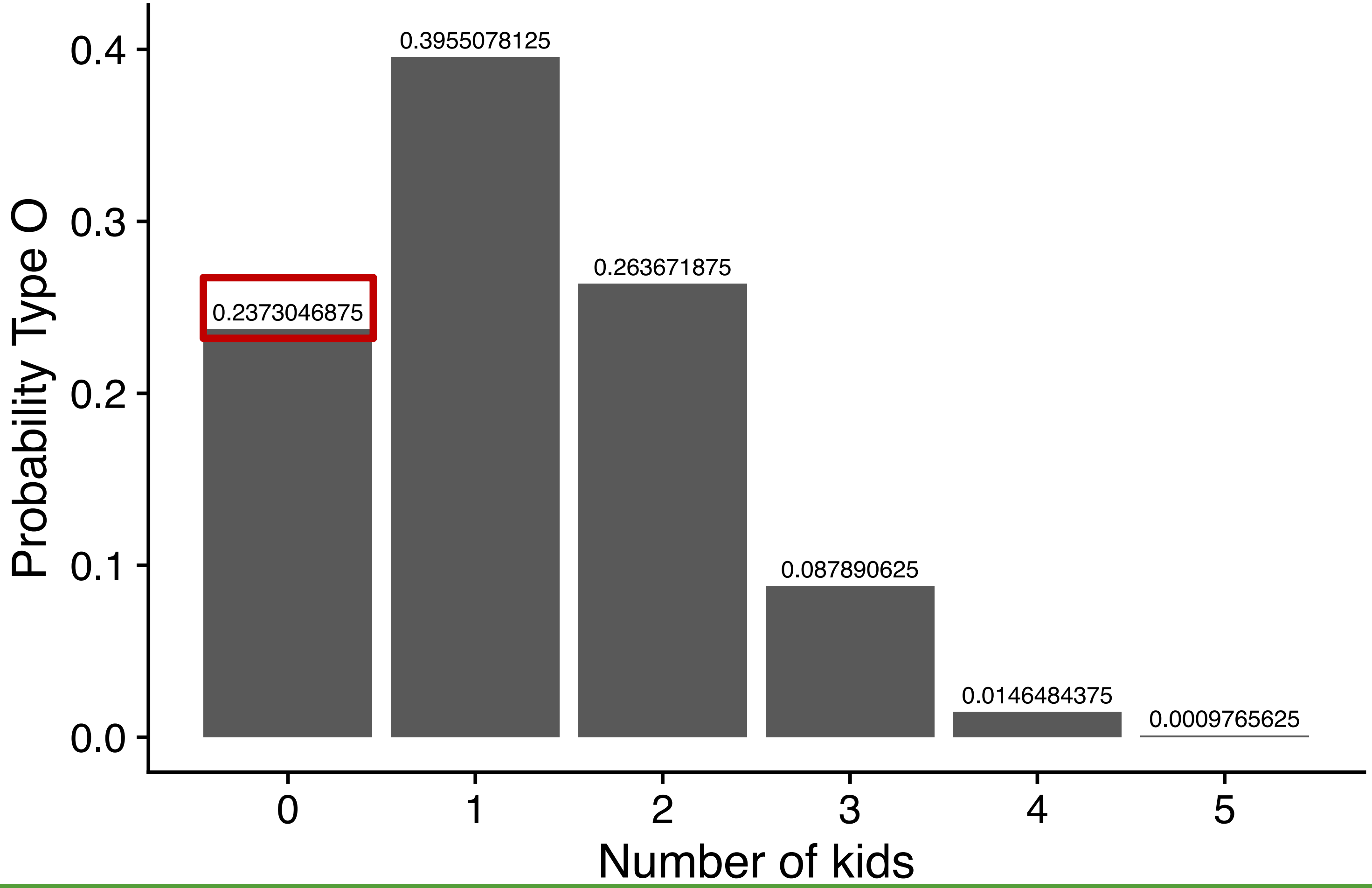
Expectation and variance

Each child born to a particular set of parents has a 25% probability of having blood type O. Assume the parents had five children. B(5, 0.25)

$\mathbb{E}(X) = \mu = \mathrm{np}$ = 5*0.25 = 1.25

$Var(X) = \sigma^2 = \mathrm{npq} = \mathrm{np}(1 - \mathrm{p})$ = 5*0.25*0.75 = 0.9375

Visualize the PMF



?distributions

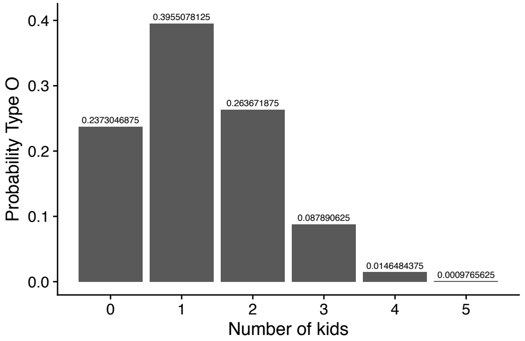
```
Distributions in the stats package

Description:
     Density, cumulative distribution function, quantile function and
     random variate generation for many standard probability
     distributions are available in the ‘stats’ package.

Details:
     The functions for the density/mass function, cumulative
     distribution function, quantile function and random variate
     generation are named in the form ‘dxxx’, ‘pxxx’, ‘qxxx’ and ‘rxxx’
     respectively.

     For the beta distribution see ‘dbeta’.

     For the binomial (including Bernoulli) distribution see ‘dbinom’.

     For the Cauchy distribution see ‘dcauchy’.

     For the chi-squared distribution see ‘dchisq’.
```

Distribution functions, generally

Function	Purpose	Binomial version
`dxxx()`	Probability distribution	`dbinom(x, size, prob)`
`pxxx()`	CDF	`pbinom(q, size, prob)`
`rxxx()`	Generate random numbers from given distribution	`rbinom(n, size, prob)`
`qxxx()`	Quantile: Inverse of pxxx()	`qbinom(p, size, prob)`

Binomial distribution functions

Binomial function	**Example**	**Output**
`dbinom(x, size, prob)`	`dbinom(2, 5, 0.25)`	Prob of obtaining 2 successes in 5 trials, where p=0.25 → **0.263**
`pbinom(q, size, prob)`	`pbinom(2, 5, 0.25)`	Prob of obtaining ≤2 successes in 5 trials, where p=0.25 → **0.896**
`rbinom(n, size, prob)`	`rbinom(100, 5, 0.25)`	Generate 100 k values from this binomial dist. → **100 from {0,1,2,3,4}**
`qbinom(p, size, prob)`	`qbinom(0.896, 5, 0.25)`	Smallest value x where F(x) >= p* → **2** *not prob success, just a prob



Making the PMF

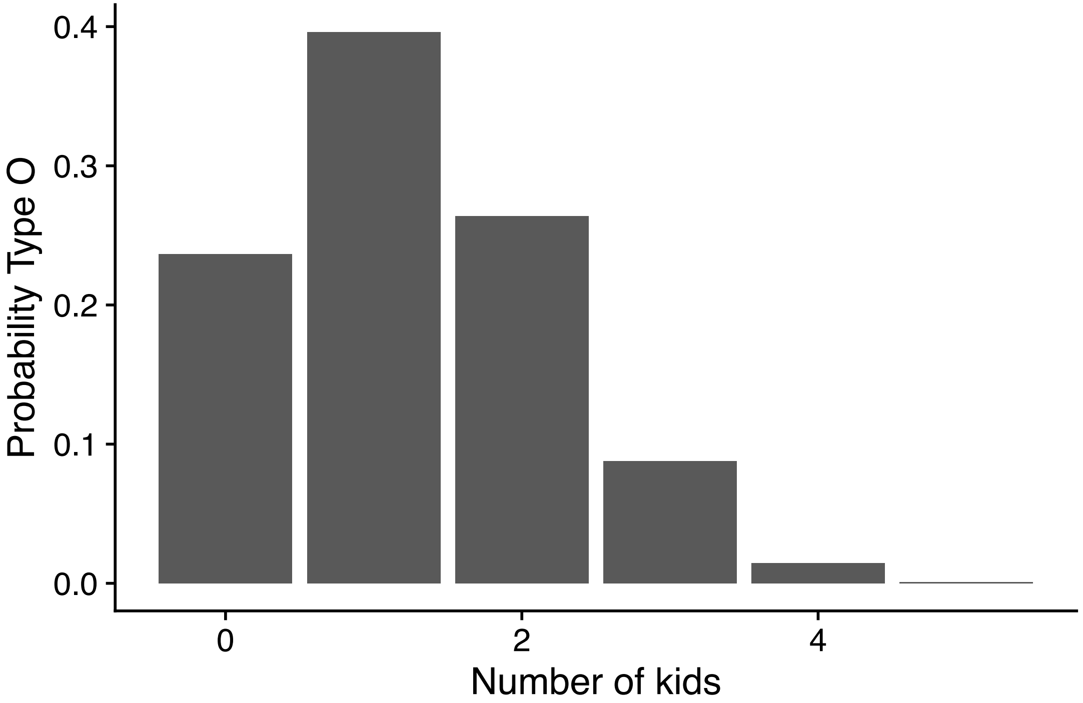
```
> ## Use dbinom() to get the PMF values
> p = 0.25
> n = 5
> k0 <- dbinom(0, 5, 0.25) ## Prob of 0 successes, aka no children are Type O
> k1 <- dbinom(1, 5, 0.25) ## Prob of 1 success, aka only 1 child is Type O

> ## Advanced:
> library(purrr)
> map_dbl(0:5, dbinom, 5, 0.25)
   [1] 0.2373046875 0.3955078125 0.2636718750 0.0878906250 0.0146484375
   [6] 0.0009765625
```

Making the PMF

```
## data frame (tibble) of probabilities for PMF
> data.pmf <- tibble(k = 0:5, prob = c(0.236623, 0.396, 0.264, 0.0879, 0.0145,
0.000977))
> data.pmf
   # A tibble: 6 x 2
         k     prob
     <int>    <dbl>
   1     0 0.236623
   2     1 0.396000
   3     2 0.264000
   4     3 0.087900
   5     4 0.014500
   6     5 0.000977

## Equivalent
> data.pmf <- tibble(k = 0:5, prob = map_dbl(0:5, dbinom, 5, 0.25))
```

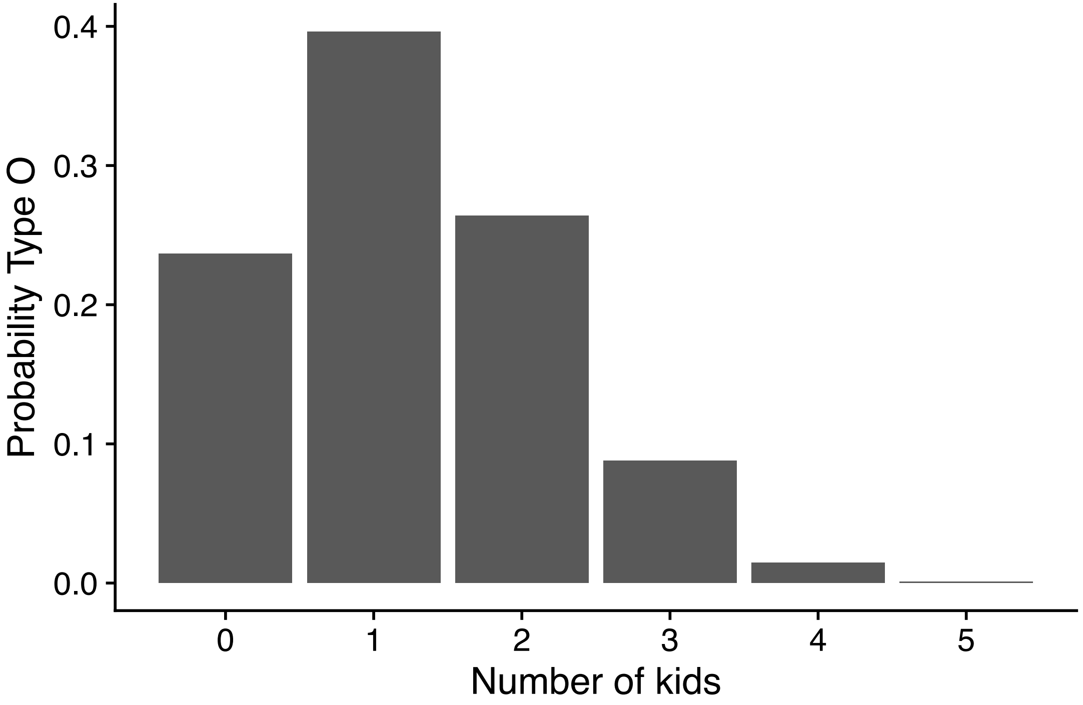
Making the PMF uses a different *stat*

```
> ggplot(data.pmf, aes(x = k, y=prob))+ geom_bar( stat="identity" ) +
         xlab("Number of kids") + ylab("Probability Type O")
```



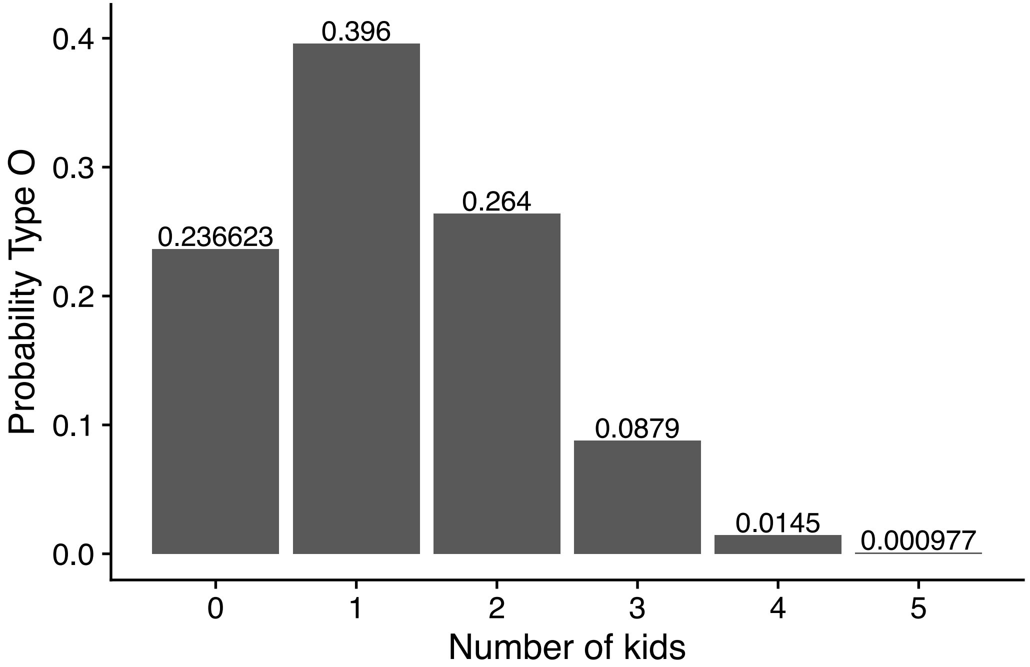
Tweaking the x-axis

```
> ggplot(data.pmf, aes(x = k, y=prob))+ geom_bar( stat="identity" ) +
          ylab("Probability Type O") +
          scale_x_continuous(name = "Number of kids", breaks = 0:5)
```



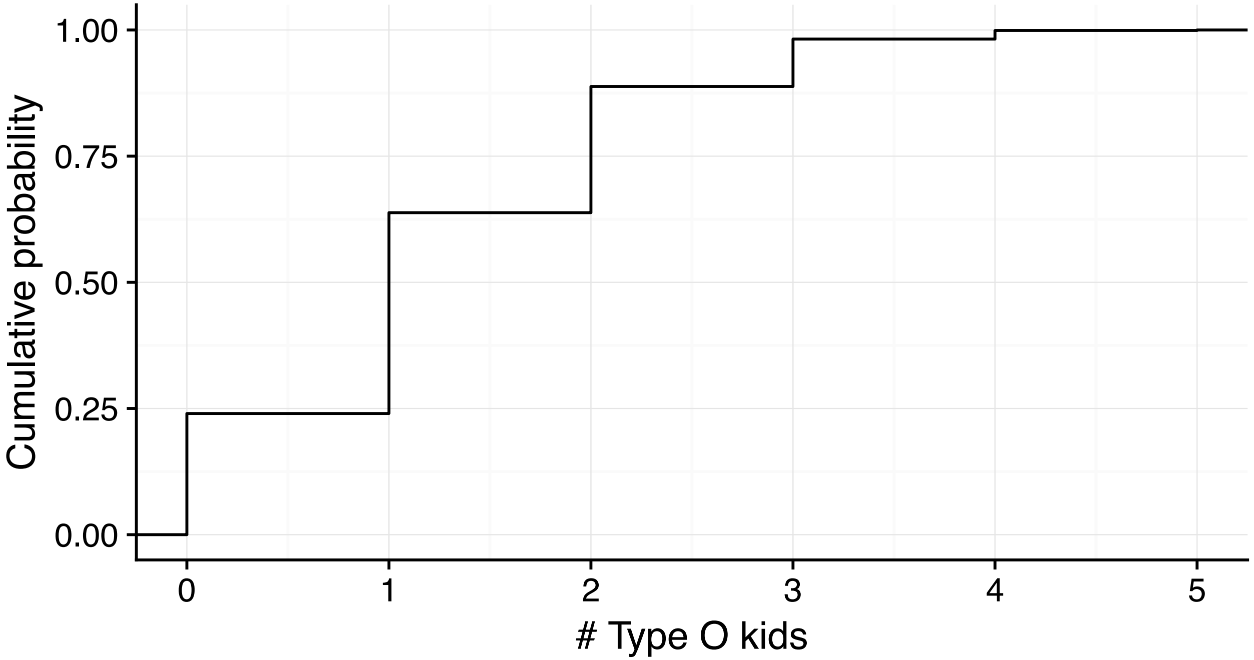
Adding some text

```
> ggplot(data.pmf, aes(x = k, y=prob))+ geom_bar( stat="identity" ) +
         ylab("Probability Type O") +
         scale_x_continuous(name = "Number of kids", breaks = 0:5) +
         geom_text(aes(x = k, y= prob + 0.01, label = prob))
```



Visualize the CDF

```
> binom.sample <- tibble(x = rbinom(1000, 5, 0.25))
> ggplot(binom.sample, aes(x=x)) + stat_ecdf() +
      xlab("# Type O kids") + ylab("Cumulative probability")
```

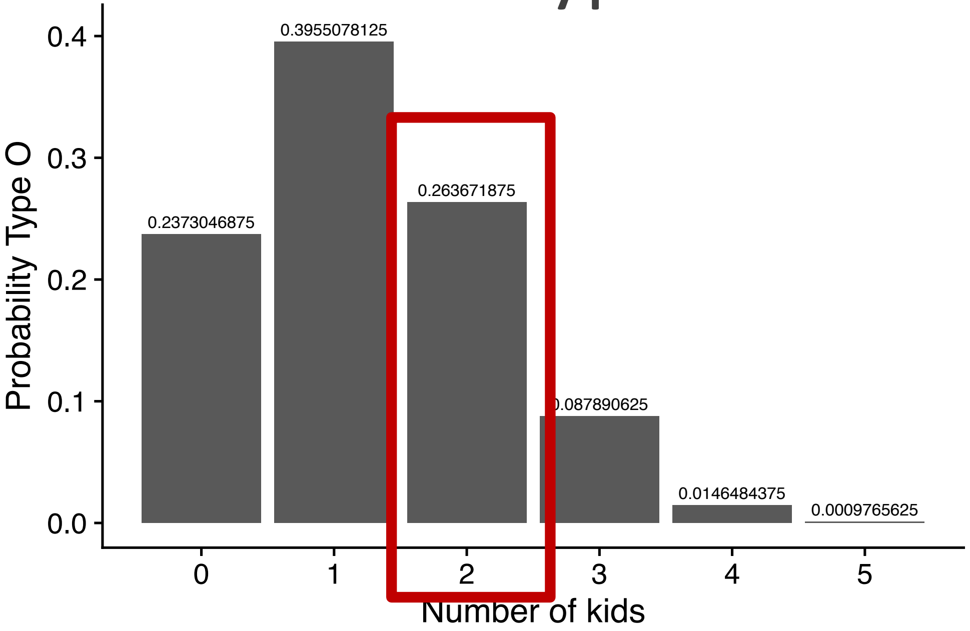


Solving for probabilities

Each child born to a particular set of parents has a 25% probability of having blood type O. Assume the parents had five children. B(5, 0.25)

What is the probability that exactly 2 children were Type O?

```
> dbinom(2, 5, 0.25)
  [1] 0.2636719
```



Solving for probabilities

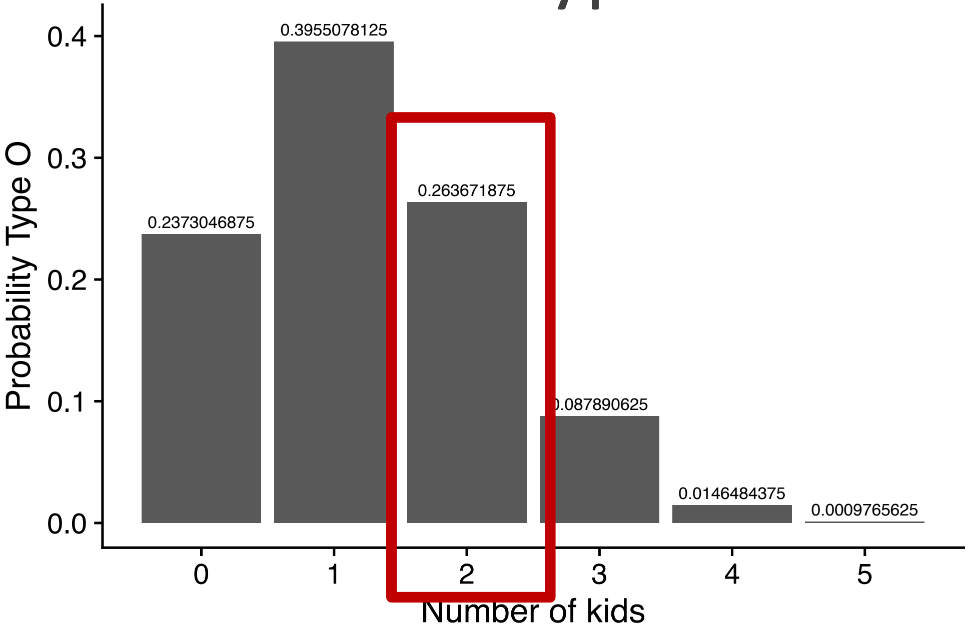
Each child born to a particular set of parents has a 25% probability of having blood type O. Assume the parents had five children. B(5, 0.25)

What is the probability that exactly 2 children were Type O?

$$P(X = k) = \binom{n}{k} p^k (1-p)^{(n-k)} = \binom{n}{k} p^k q^{(n-k)}$$

$$P(X = 2) \;= \binom{5}{2} 0.25^2 0.75^{(5-2)}$$

= 10 * 0.0625 * 0.422 = 0.26375

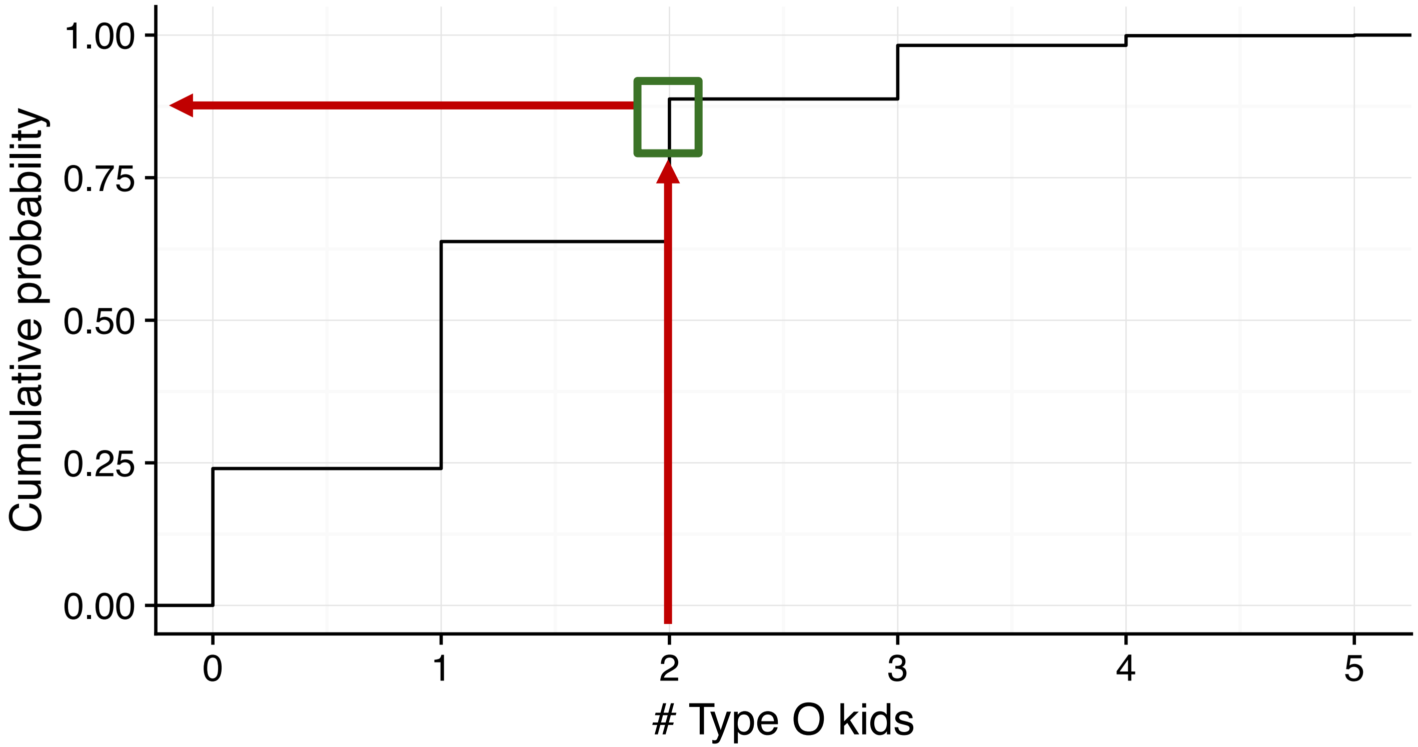


Solving for probabilities

B(5, 0.25)

What is the probability that 2 or fewer children were Type O?

```
> pbinom(2, 5, 0.25)
  [1] 0.8964844
```

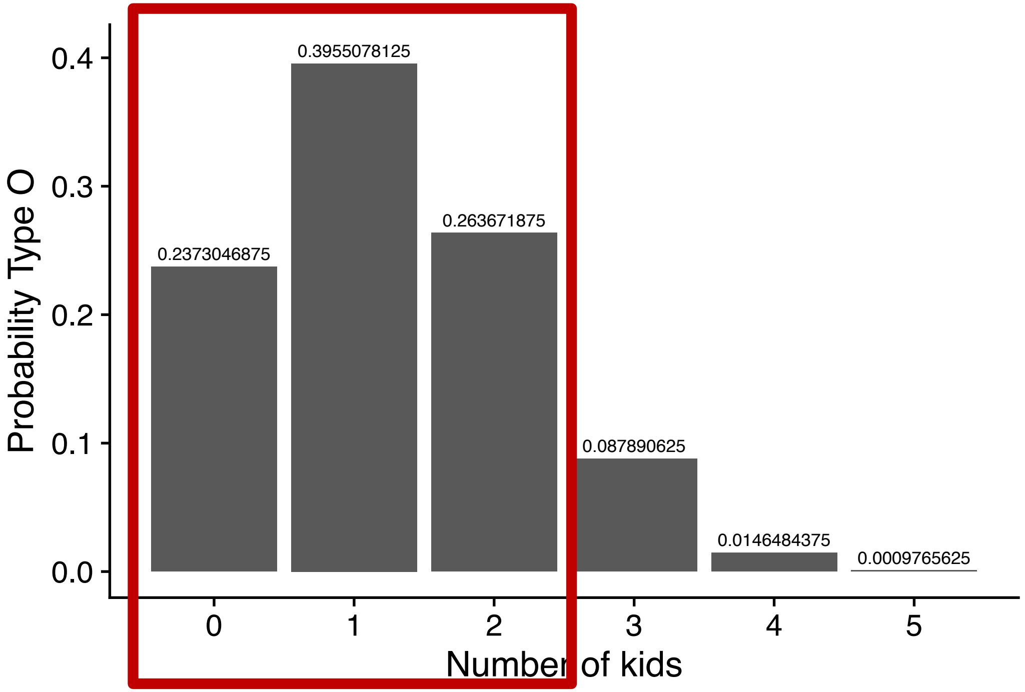


Solving for probabilities

B(5, 0.25)

What is the probability that *2* or fewer children were Type O?

```
> dbinom(0, 5, 0.25) +
  dbinom(1, 5, 0.25) +
  dbinom(2, 5, 0.25)

[1] 0.8964844
```

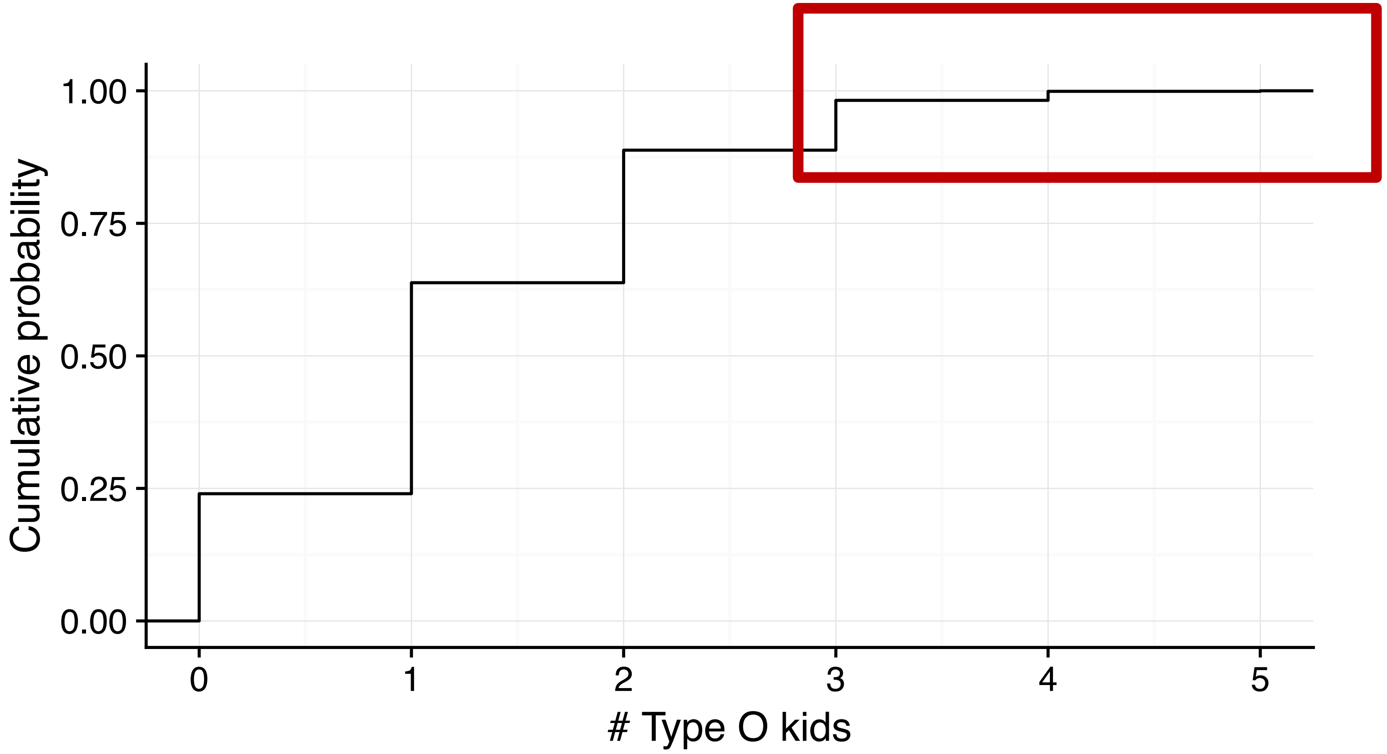


Solving for probabilities

B(5, 0.25)

What is the probability that *more than 2* children (ie either 3, 4, or 5) were Type O?

```
> 1 - pbinom(2, 5, 0.25)
   [1] 0.1035156
```

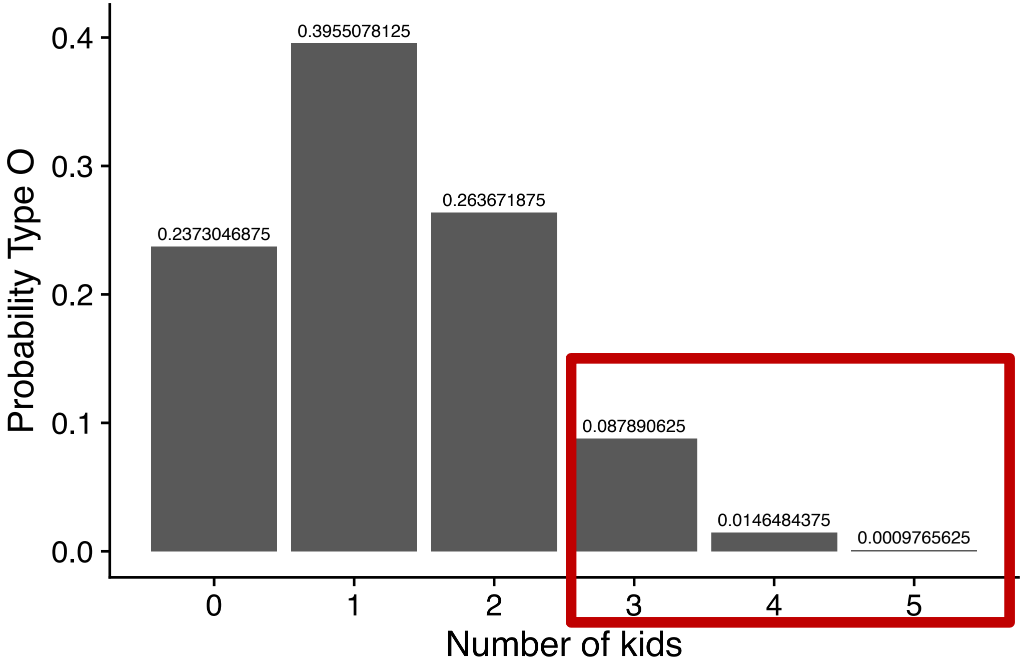


Solving for probabilities

B(5, 0.25)

What is the probability that *more than 2* children (ie either 3, 4, or 5) were Type O?

```
> dbinom(3, 5, 0.25) +
  dbinom(4, 5, 0.25) +
  dbinom(5, 5, 0.25)

  [1] 0.1035156
```



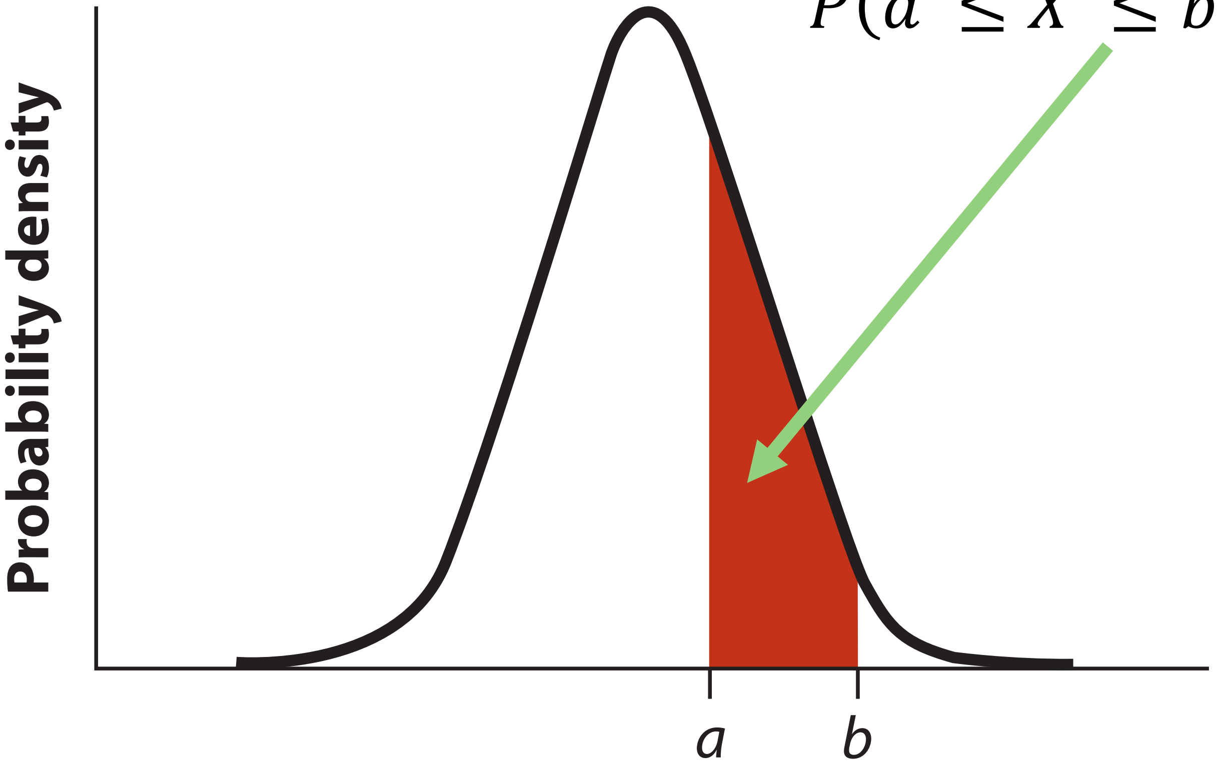
BREATHE

Expressing continuous random variables

Probability density function (PDF)

- Describes the values taken by a continuous r.v. X and its associated probabilities
- Function such that the area under the curve between any two points a, b corresponds to the probability that the r.v. falls between a, b
- → $P(a \leq X \leq b) = \int_a^b f(x)dx$

PDF



$$P(a \ \leq X \ \leq b) = \int_{a}^{b} f(x)dx$$

PDF properties

Continuous r.v.'s are *infinitely precise*: $P(X = x) = P(x \leq X \leq x) = 0$

- **Exactly unlike PMFs**

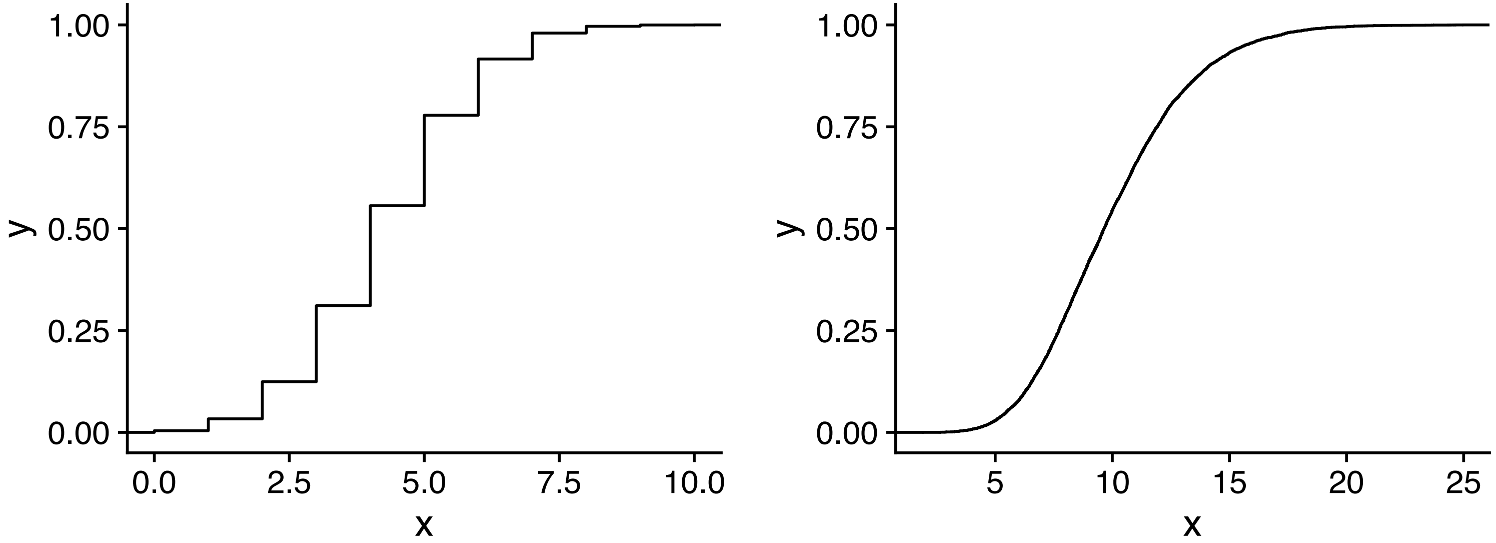
Total area under the PDF equals 1: $\int_{-\infty}^{\infty} f(x)dx = 1$

Probabilities aren't negative: $f(x) \geq 0$

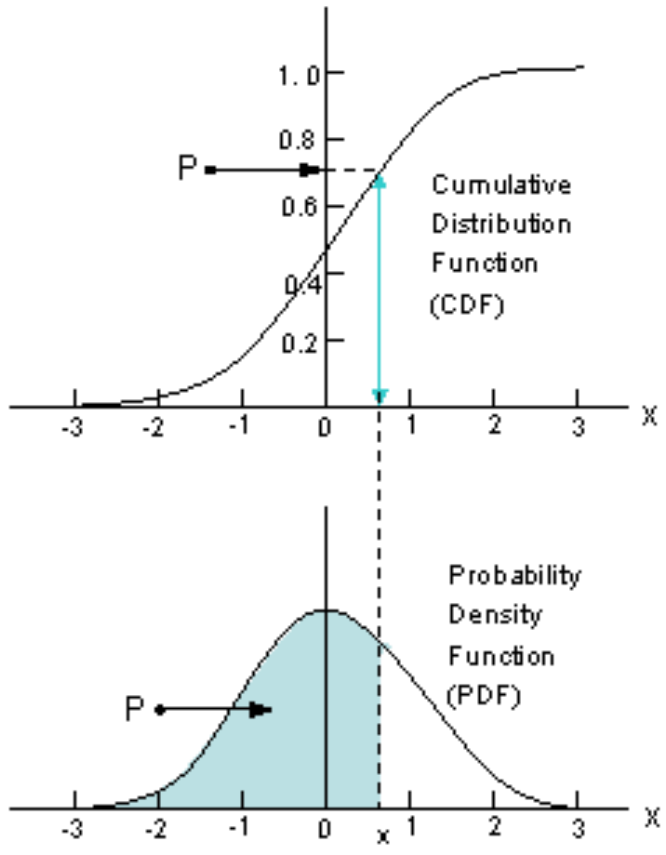
Expressing continuous random variables

Cumulative distribution function (CDF)

- Function defined, for a specific value x of a continuous r.v. $\mathcal{X}$, as $\mathcal{F}(x) = \mathcal{P}(\mathcal{X} \leq x)$
- (mostly) the same as for discrete



Relationship between PDF and CDF



Jumping right in: Normal distribution

The PDF (probability distribution) for a normally-distributed random variable:

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

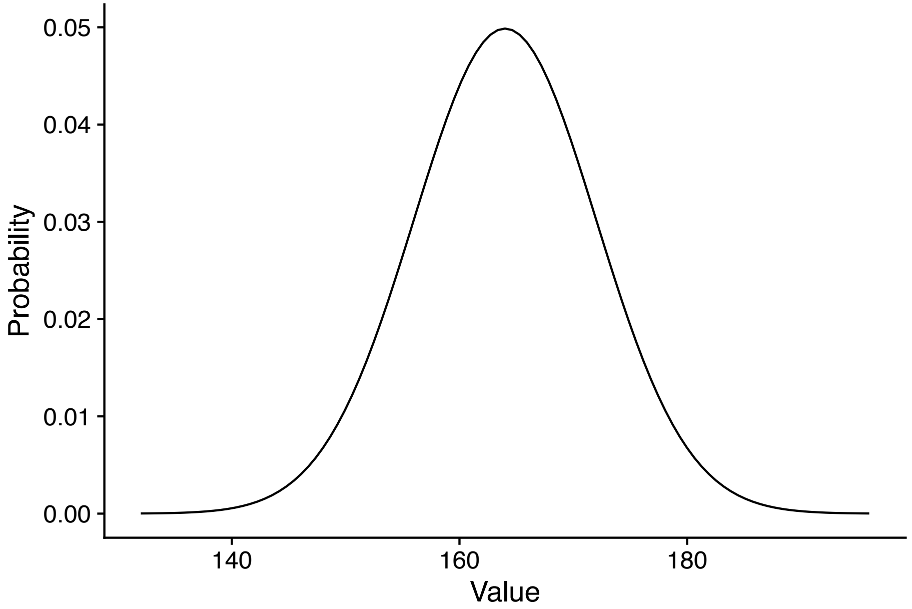
It's gross, everyone knows it, and you will be neither plugging nor chugging with this equation

We write normally distributed r.v.'s as $X \sim N(\mu, \sigma^2)$

PDF of normal distribution

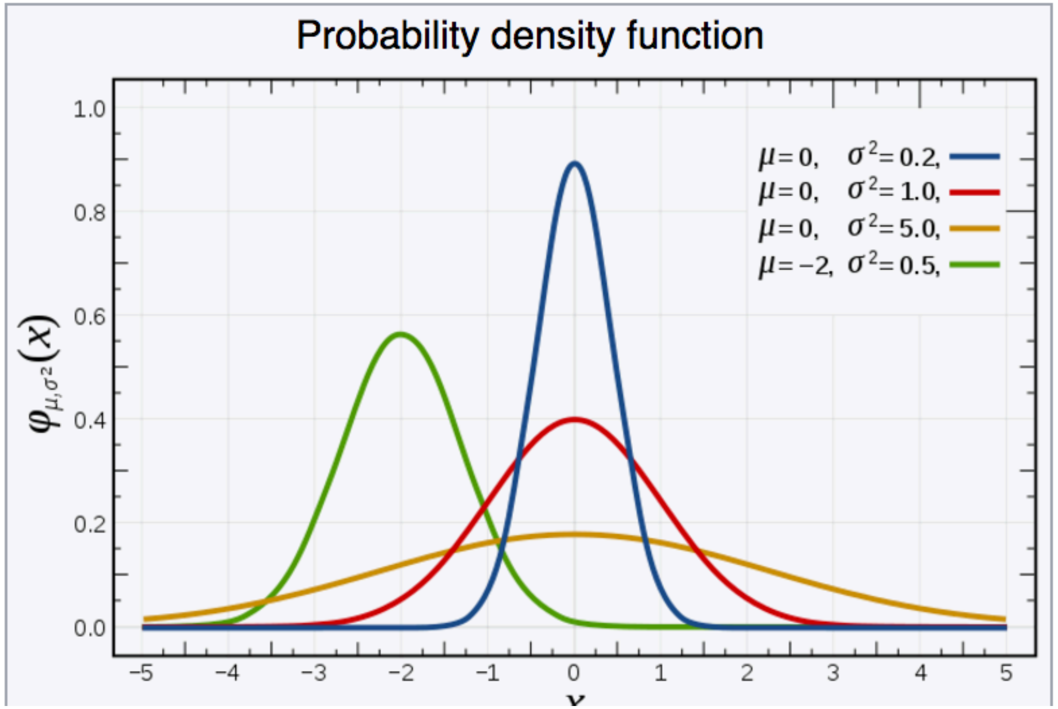
Example, let's say women's heights (cm) are normally distributed according to $N(165, 64)$

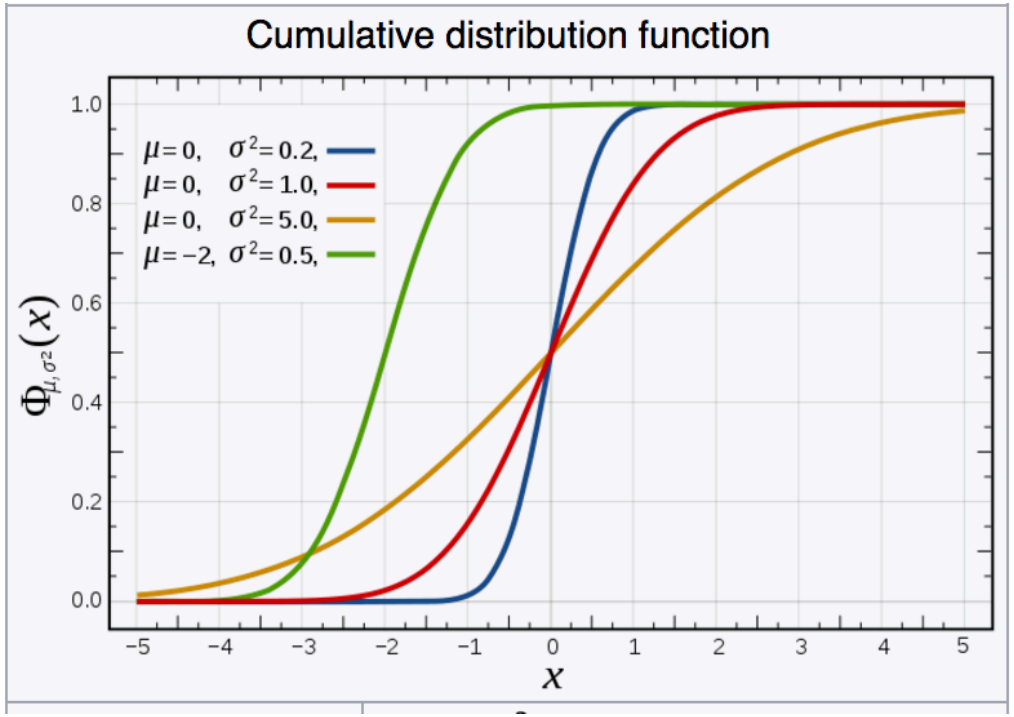
- Pop quiz: what is the standard deviation of this distribution?



Wikipedia weighs in

Normal distribution

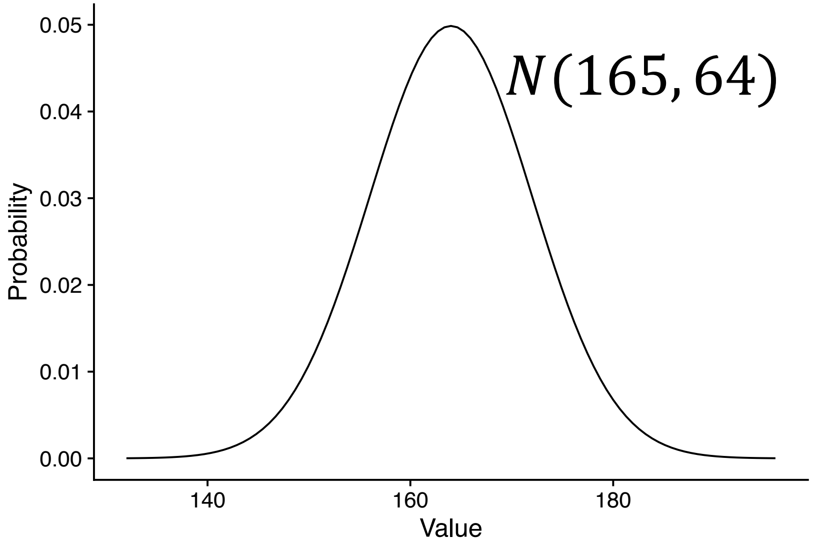




Making the PDF

Another "interesting" hack:

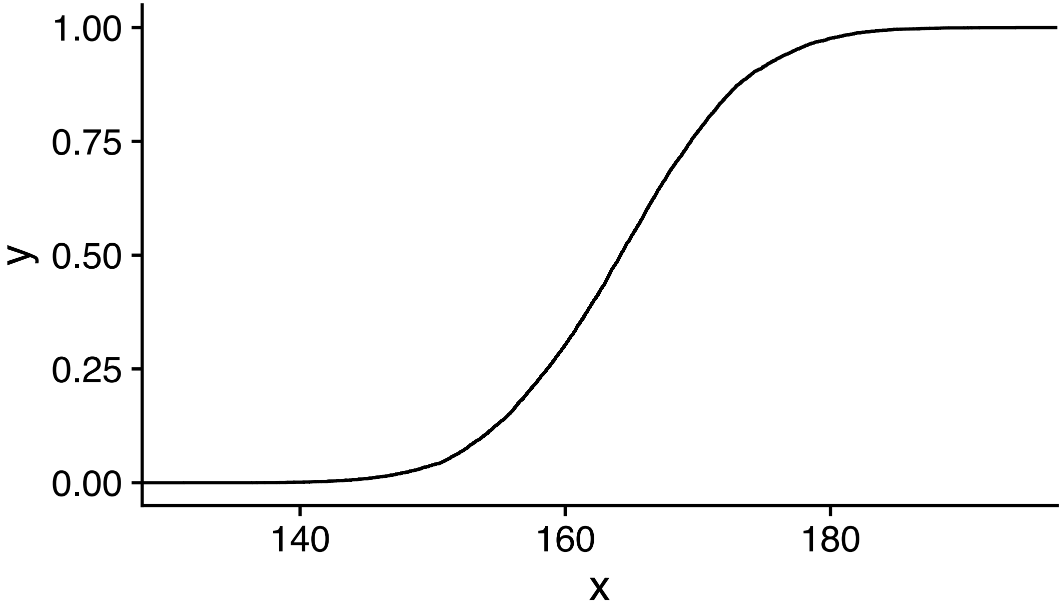
```
> plot.range <- tibble(x = c(165 - 32, 165 + 32))
> ggplot(plot.range, aes(x=x)) +
      stat_function(fun = dnorm,  args=list(mean=164, sd=8))
```



$N(165, 64)$

Making the CDF

```
> data.cdf <- tibble(x = rnorm(10000, 164, 8))
> ggplot(data.cdf, aes(x=x)) + stat_ecdf()
```



Expectation and variance

Any guesses?

It's in the definition: $X \sim N(\mu, \sigma^2)$

Working with the normal distribution

Types of questions one can ask:

- What is the probability that a randomly-chosen woman is taller than 158 cm?
- What is the probability that a randomly-chosen woman is between 163—170 cm tall?
- What is the probability that a randomly-chosen woman is shorter than 167 cm?
- What is the probability that a randomly-chosen woman is 168 cm tall?

Working with the normal distribution

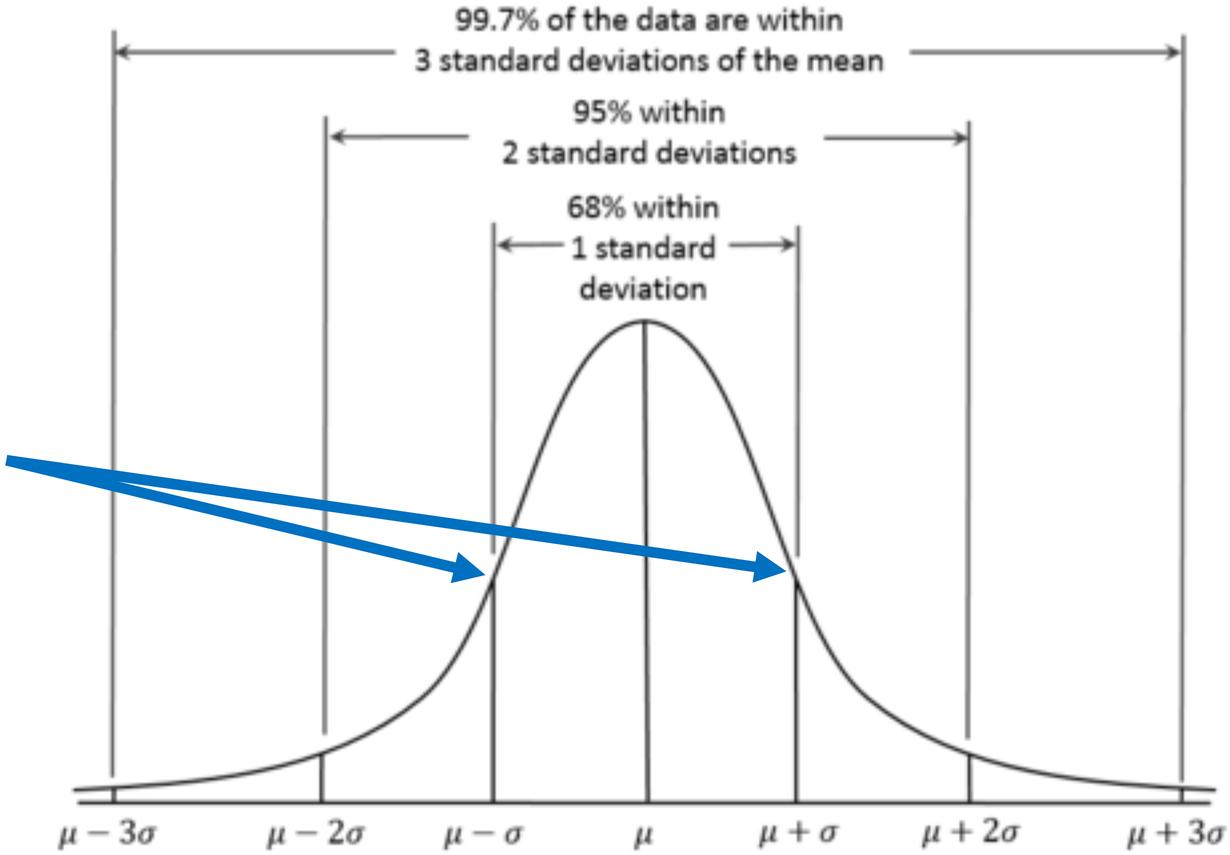
Types of questions one can ask:

- What is the probability that a randomly-chosen woman is taller than 158 cm?
- What is the probability that a randomly-chosen woman is between 163—170 cm tall?
- What is the probability that a randomly-chosen woman is shorter than 167 cm?
- ~~What is the probability that a randomly-chosen woman is 168 cm tall?~~

Properties of the normal distribution

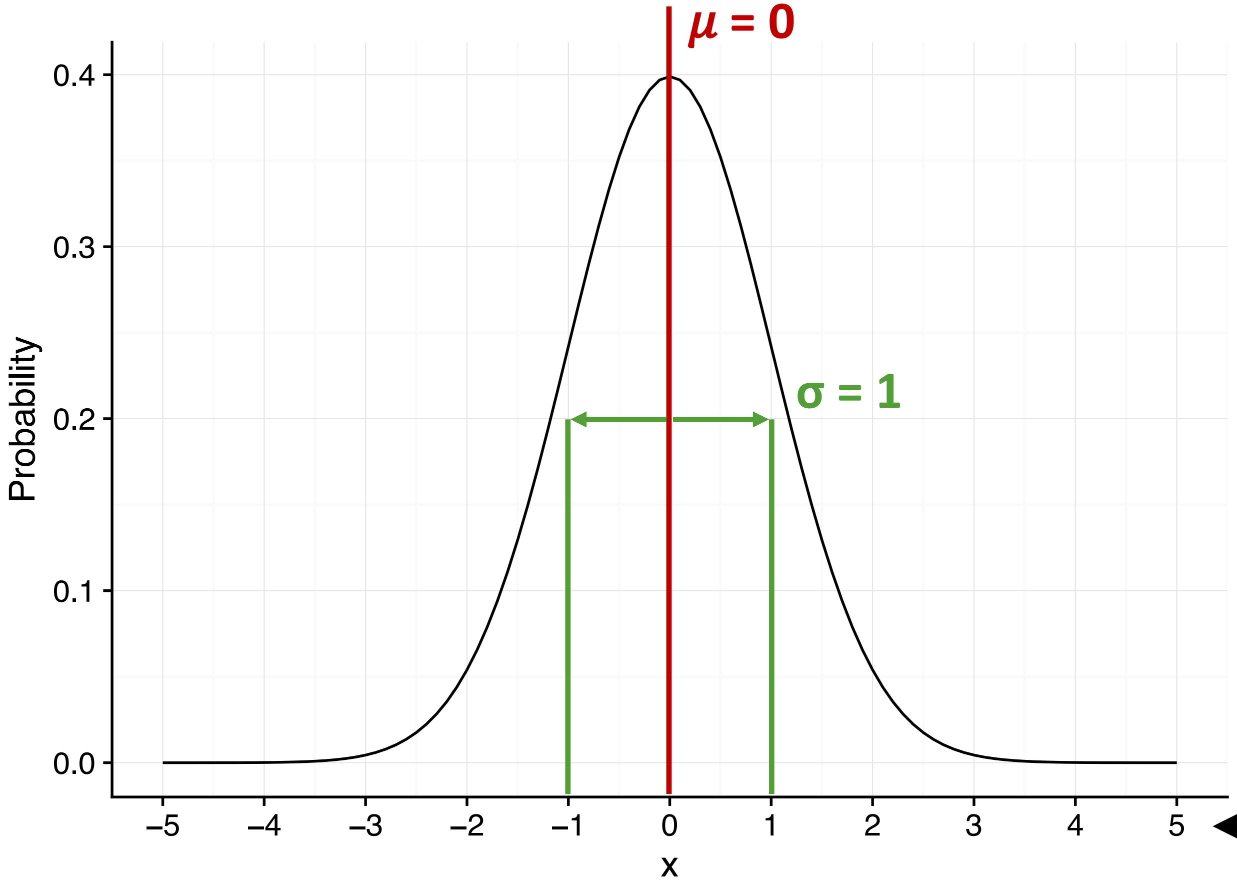
Symmetric around the mean

Mean = median = mode


99.7% of the data are within 3 standard deviations of the mean
95% within 2 standard deviations
68% within 1 standard deviation
$\mu - 3\sigma$
$\mu - 2\sigma$
$\mu - \sigma$
μ
$\mu + \sigma$
$\mu + 2\sigma$
$\mu + 3\sigma$

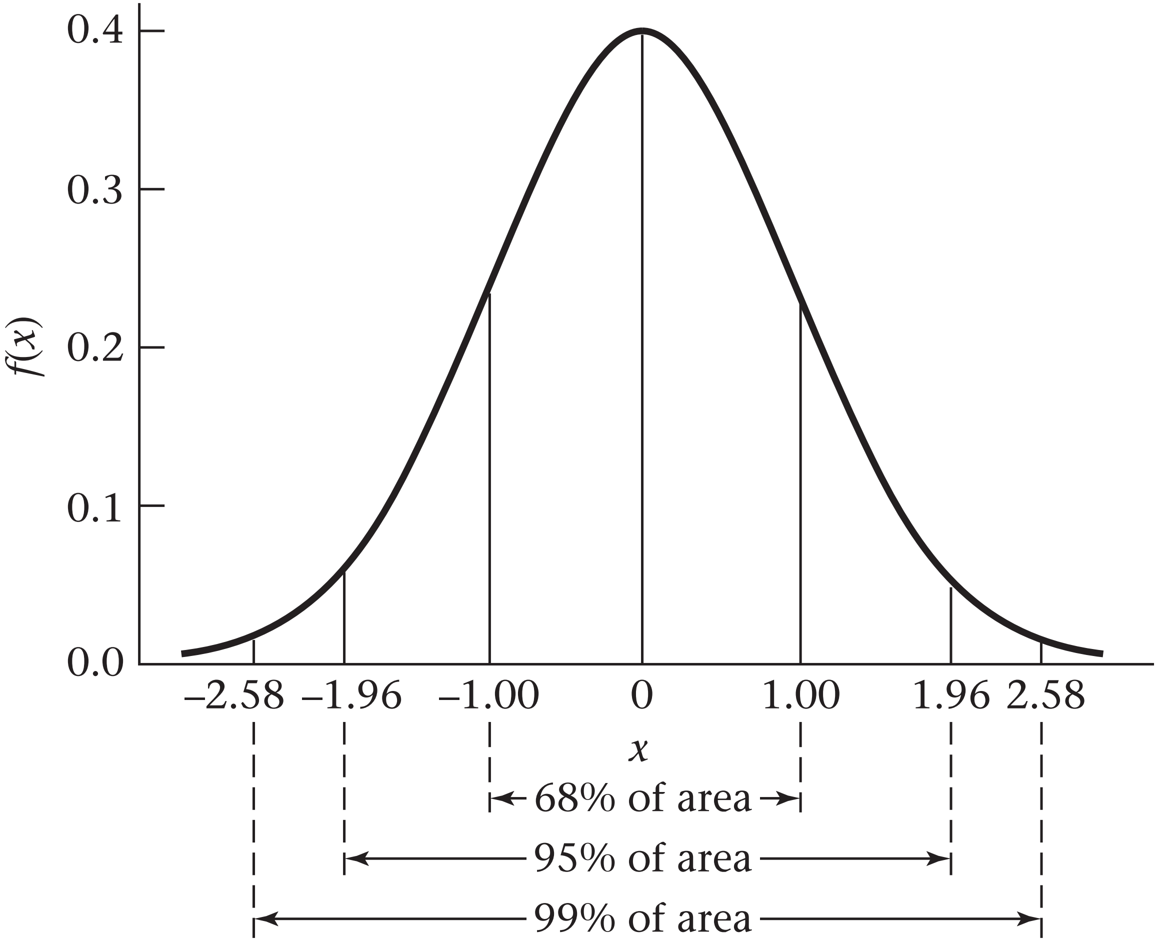
Inflection points

Introducing the **standard normal**: $X \sim N(0,1)$



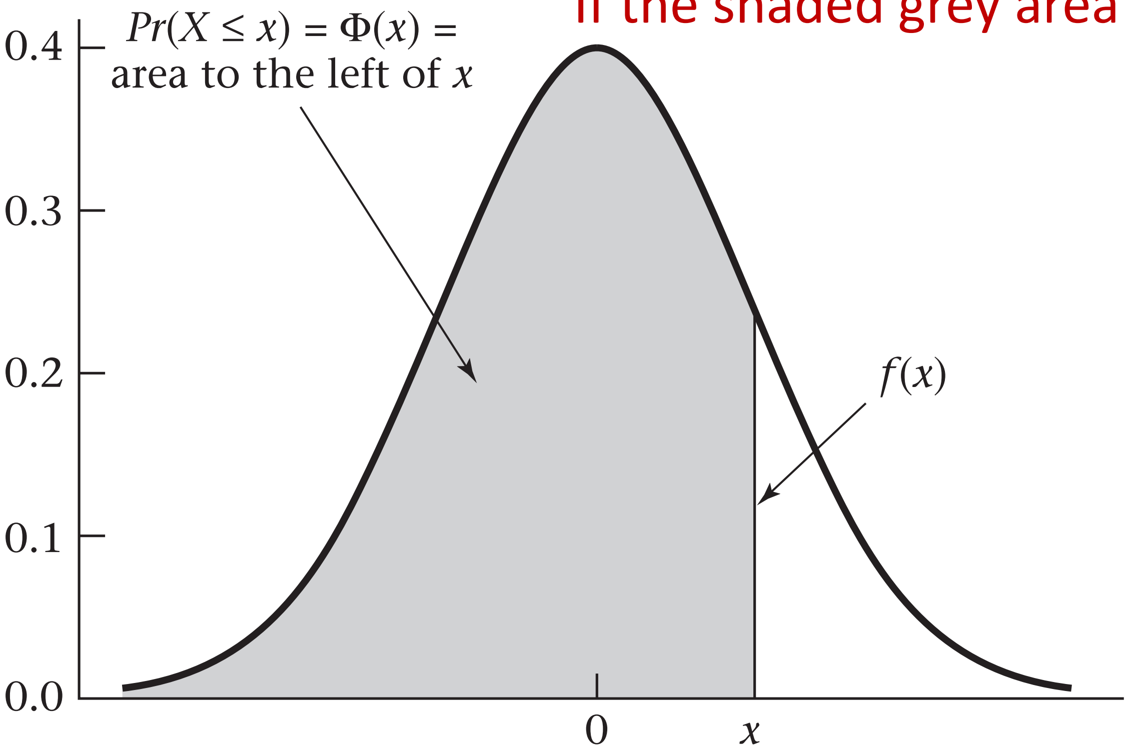
These are **Z-scores**

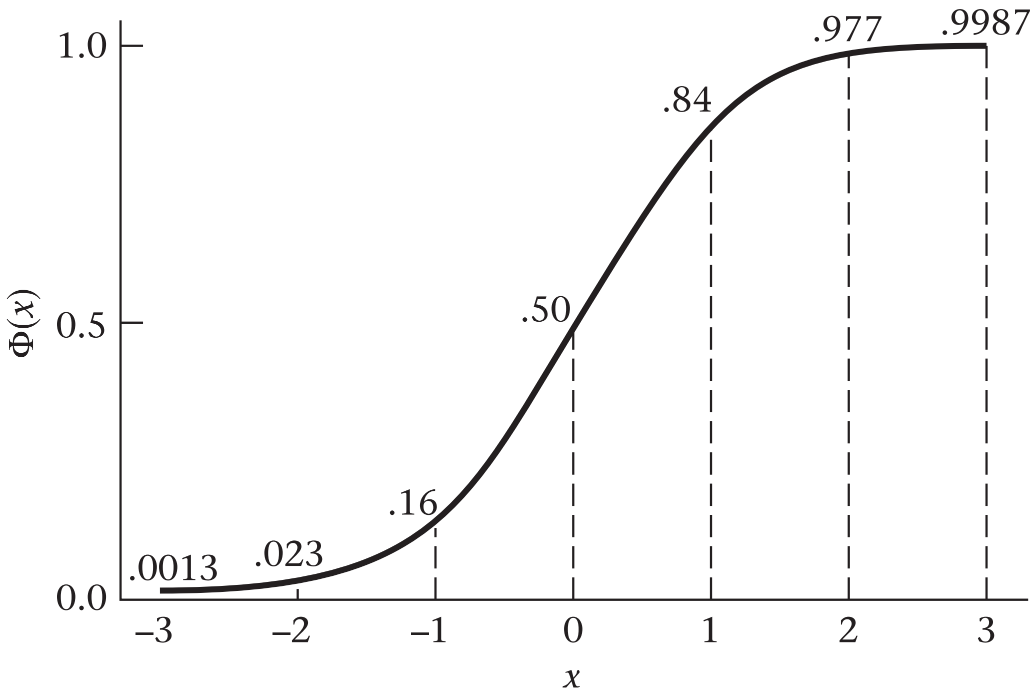
Standard Normal $X \sim N(0,1)$



PDF and CDF of $X \sim N(0,1)$

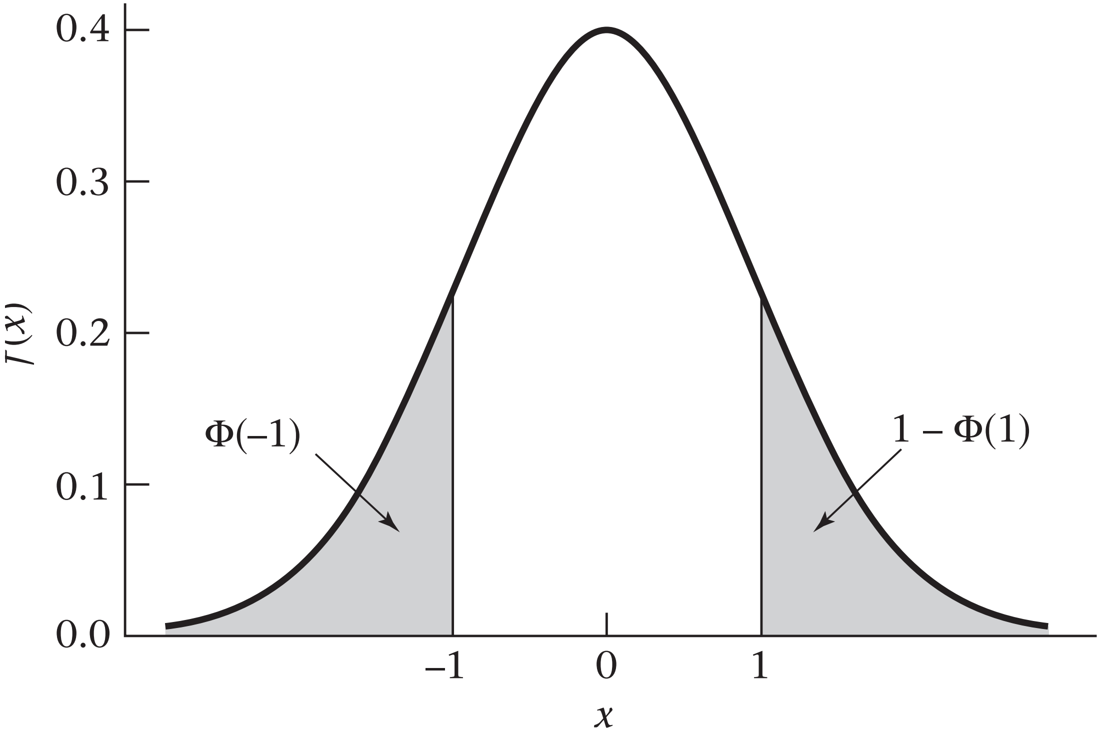
If the shaded grey area = 0.977, what is x?





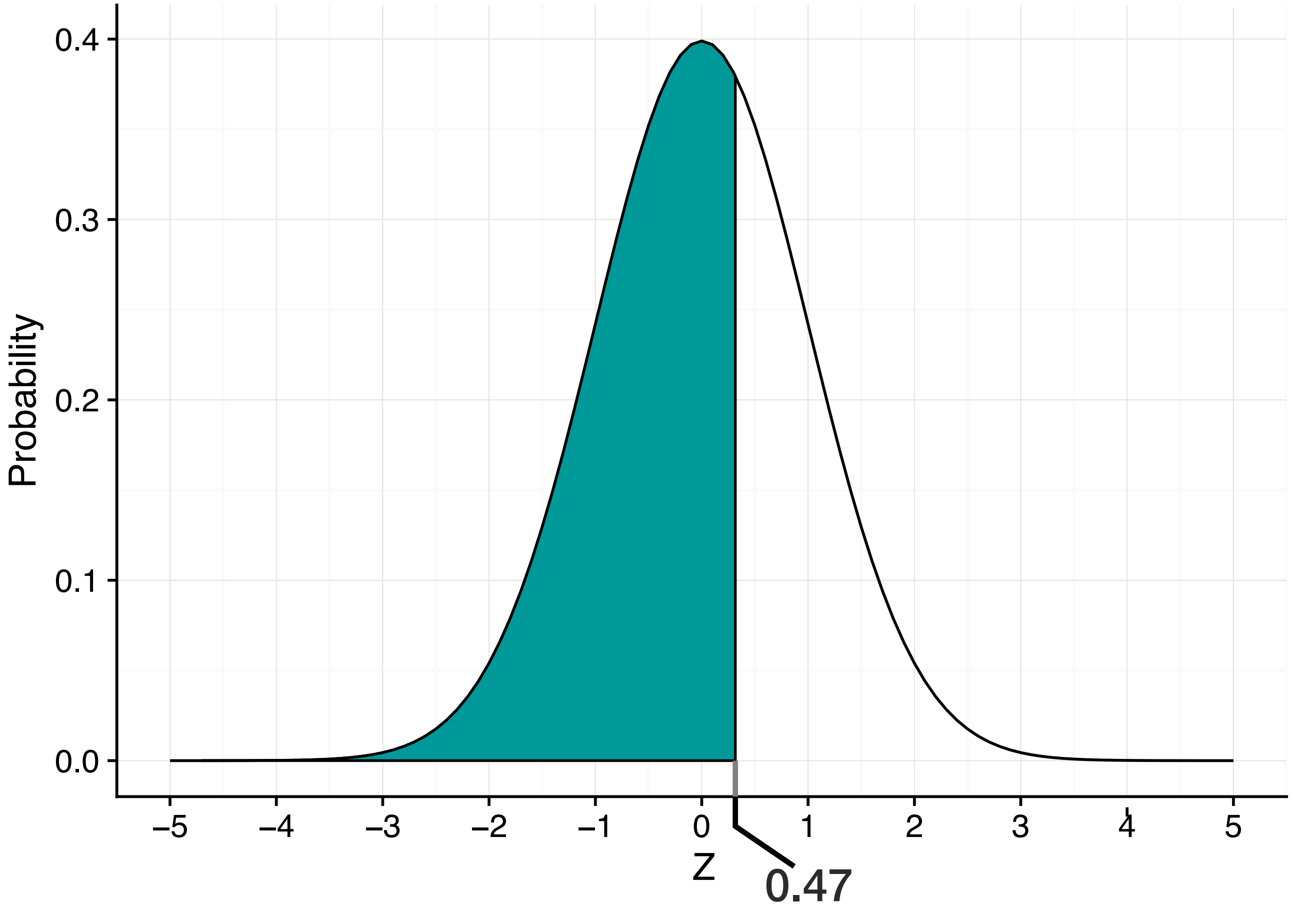
Standard Normal $X \sim N(0,1)$

Due to symmetry, P(X ≤ -x) = 1 - P(X ≤ x)



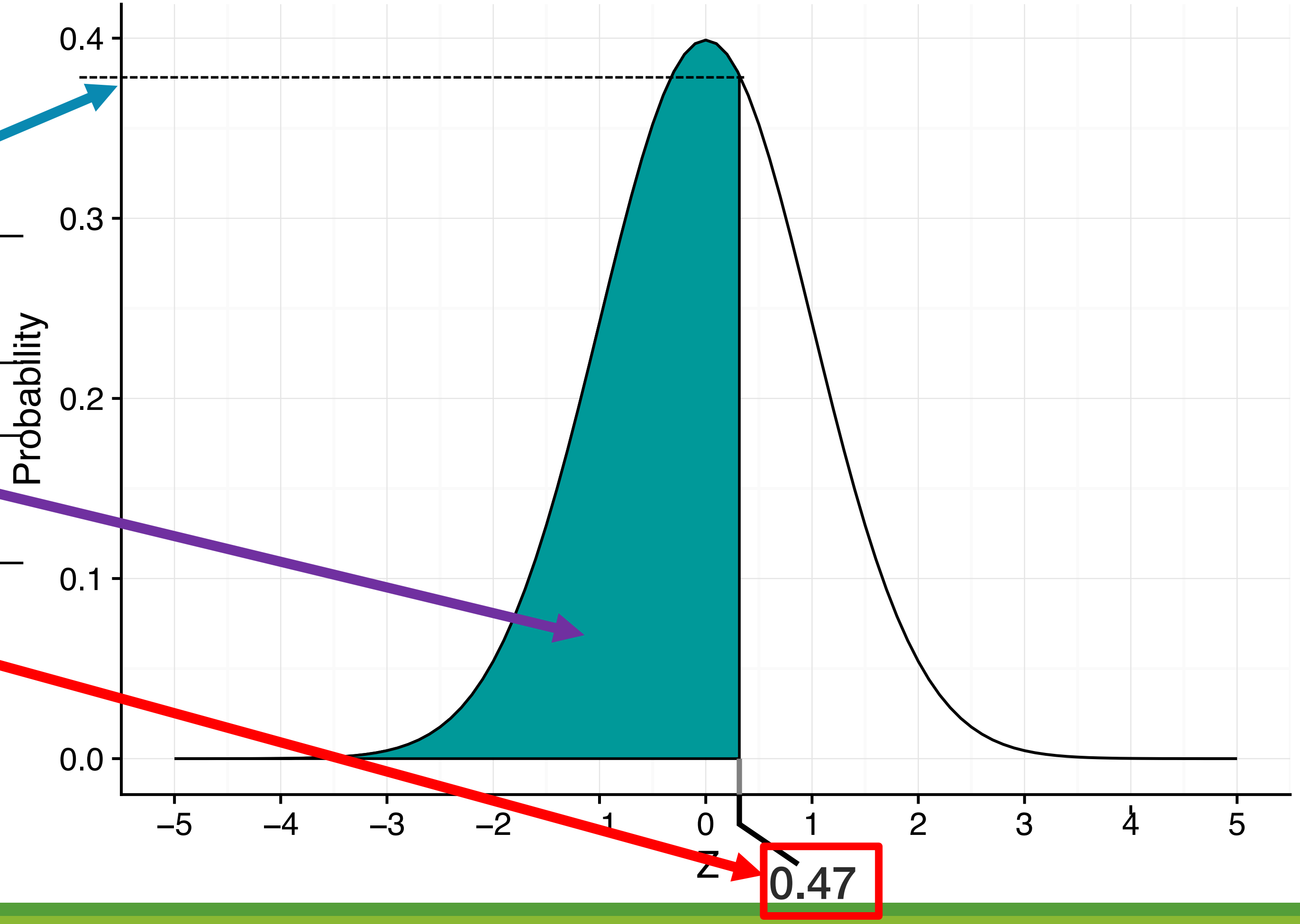
For $X \sim N(0,1)$, what is the probability P(X ≤ 0.47)?

```
# CDF: P(X <= 0.47)
> pnorm(0.47)
   [1]  0.6808225
```

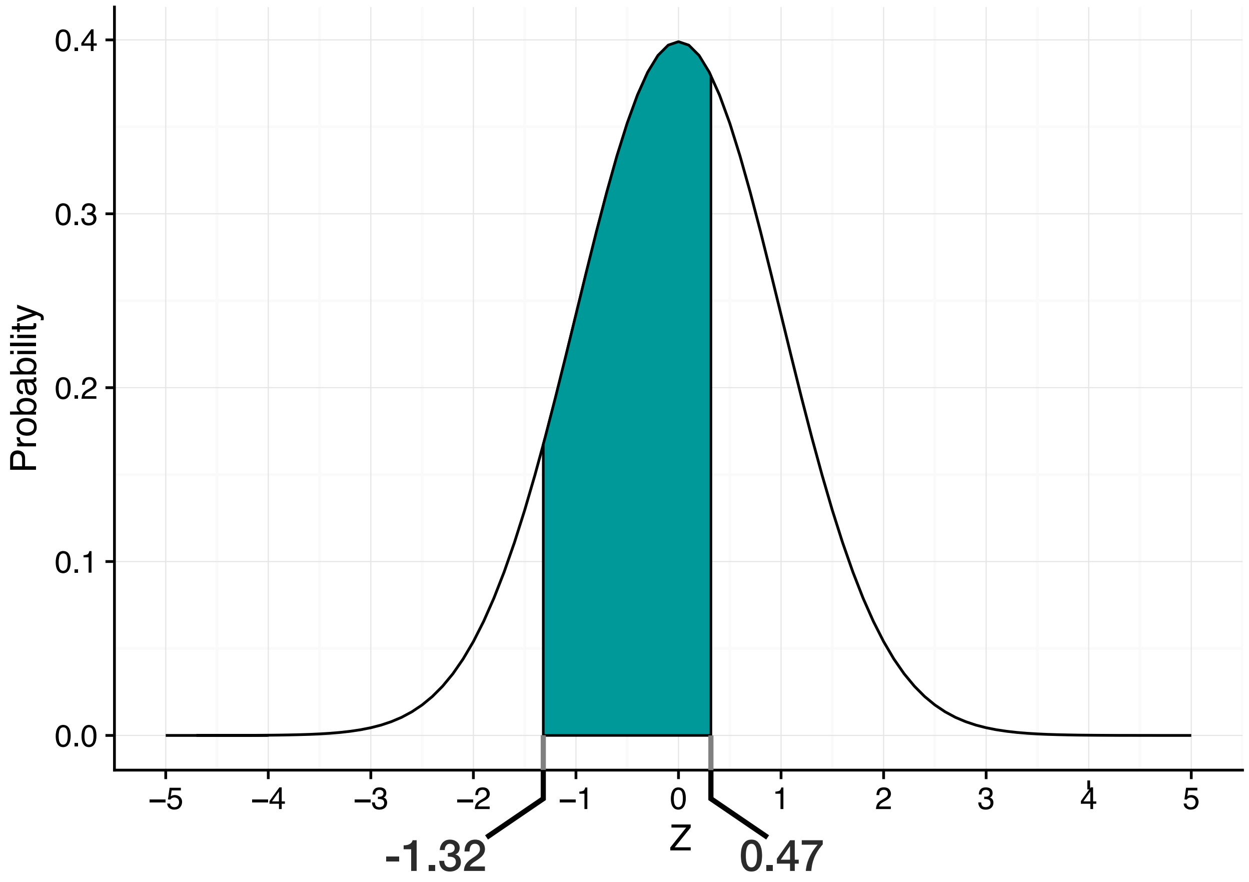


Normal distribution functions

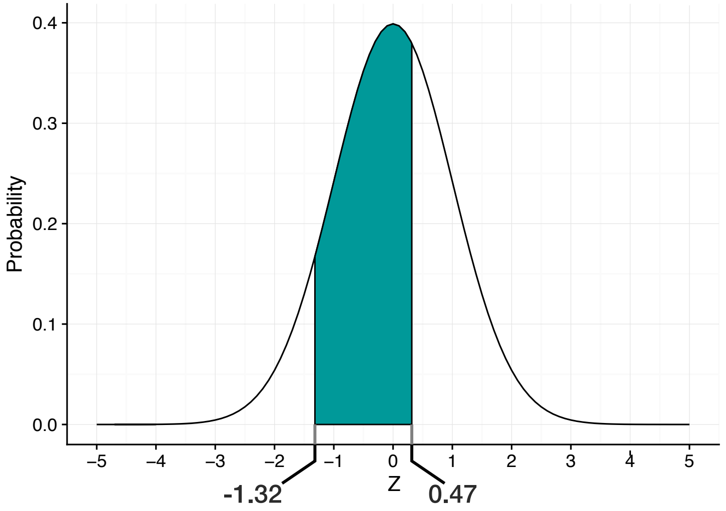
Normal function	Meaning
dnorm(x)	Density at X=x
pnorm(q)	P(X <= x)
rnorm(n)	Generate n random draws from N(0,1)
qnorm(p)	Obtain x from given CDF area: `qnorm(0.6808225) = 0.47`



For $X \sim N(0,1)$, what is the probability P(-1.32 ≤ x ≤ 0.47)?

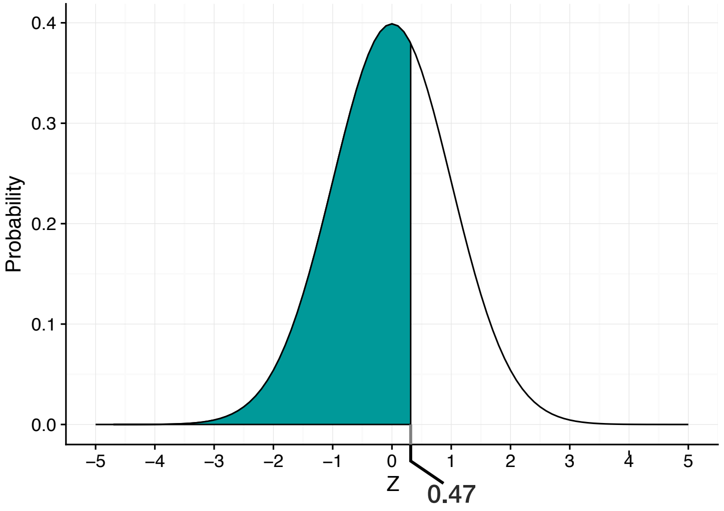


For $X \sim N(0,1)$, what is the probability P(-1.32 ≤ x ≤ 0.47)?



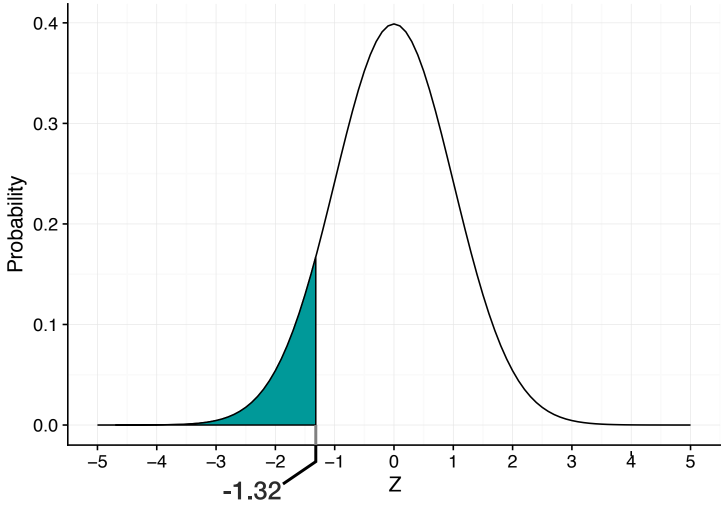
0.587405

=



```
# P(X <= 0.47)
> pnorm(0.47)
   [1]  0.6808225
```

-



```
# P(X <= -1.32)
> pnorm(-1.32)
   [1] 0.09341751
```

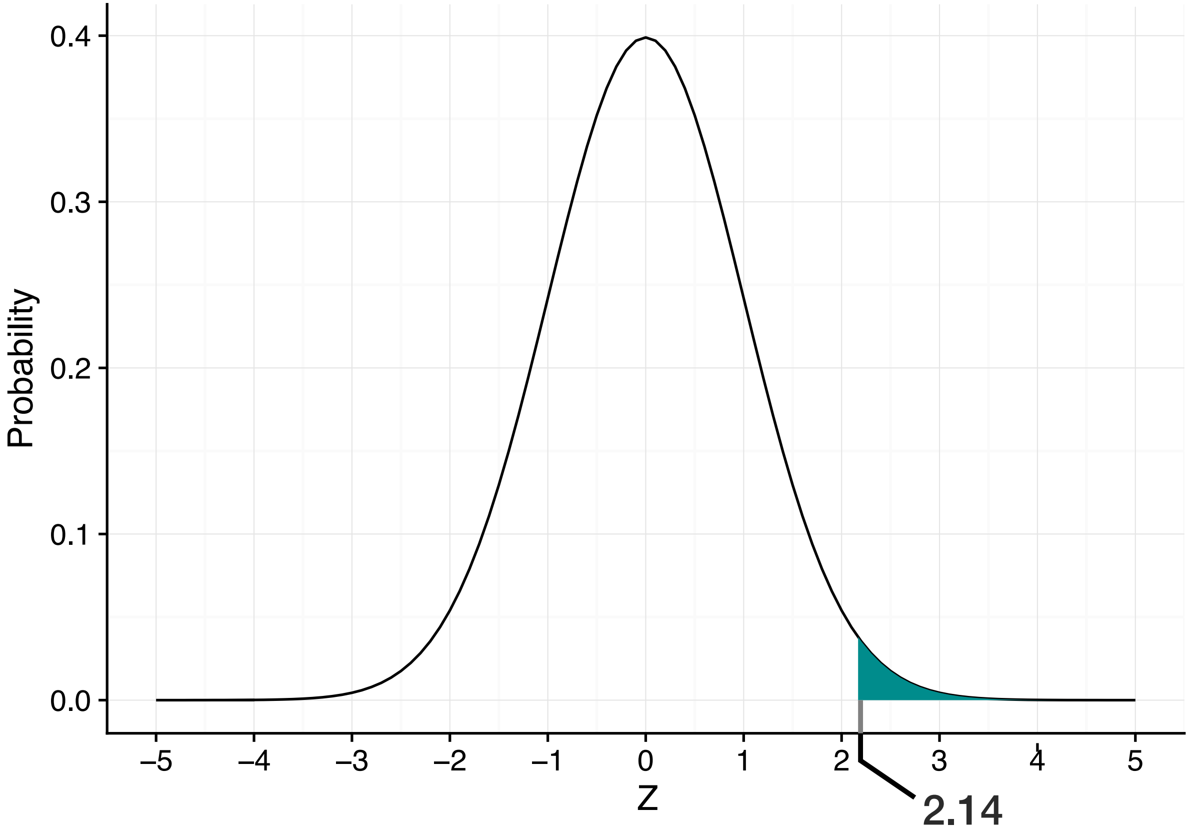
For $X{\sim}N(0,1)$, what is the probability P(-1≤ x ≤ 1)?

AKA probability of being within 1 standard deviation of mean?

~0.68

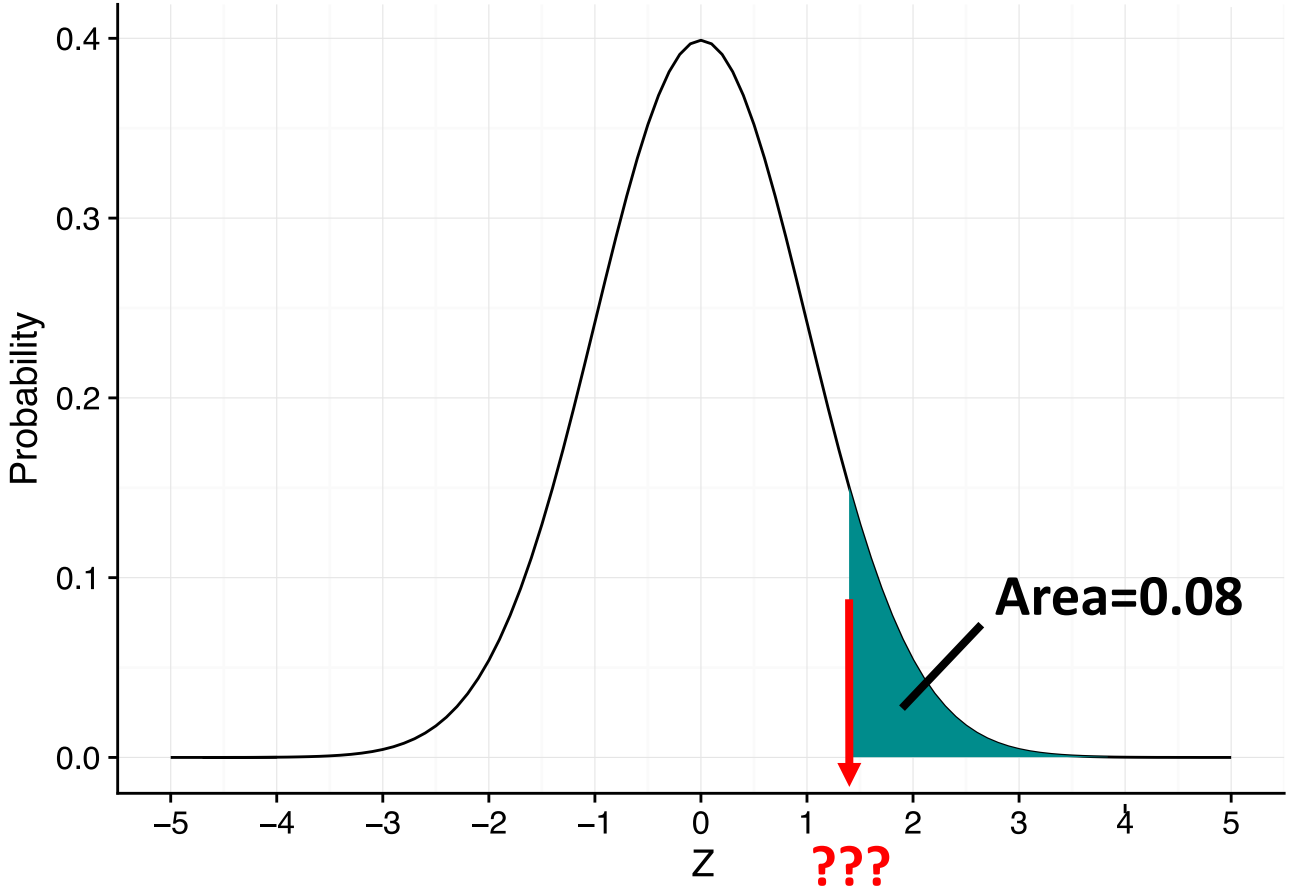
For $X \sim N(0,1)$, what is the probability P(x ≥ 2.14)?

```
## Two approaches:

> 1 - pnorm(2.14)
   [1] 0.01617738

> pnorm(-2.14)
   [1] 0.01617738
```



For $X \sim N(0,1)$, the top 8% of the distribution falls above what number?

```
> qnorm(1 - 0.08)
   [1] 1.405072

> -1 * qnorm(0.08)
   [1] 1.405072
```



Historical consideration of z-scores

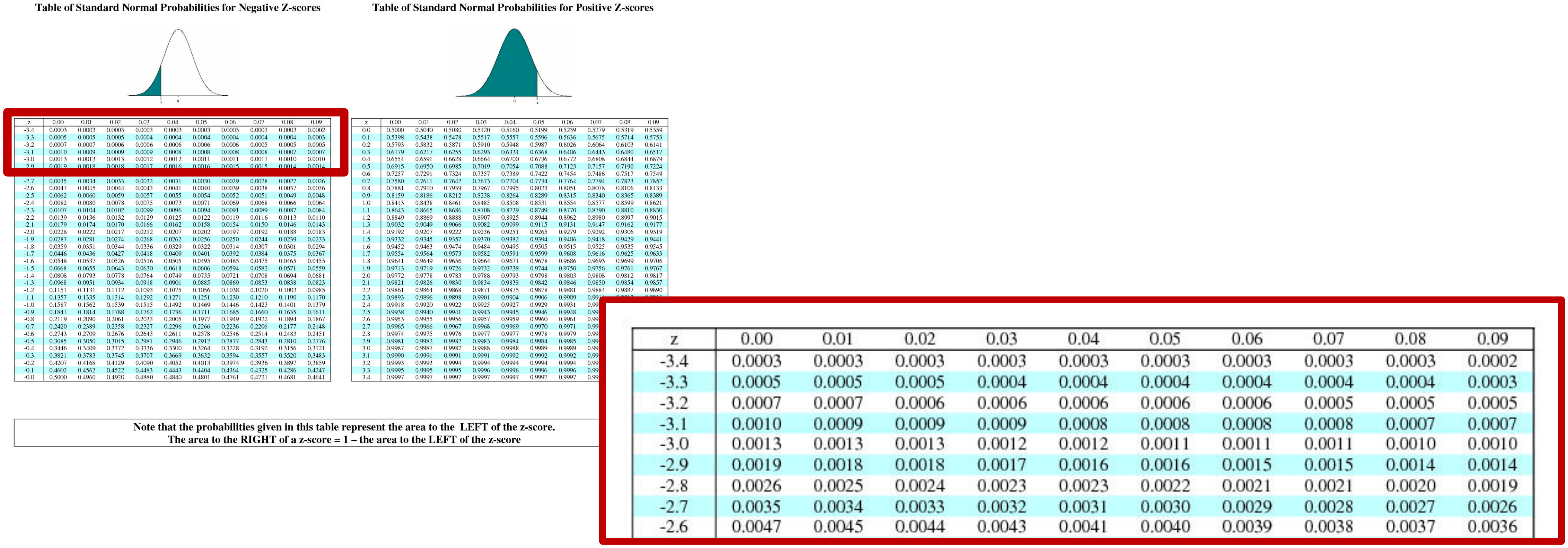


Table of Standard Normal Probabilities for Negative Z-scores

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
-3.4	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0002
-3.3	0.0005	0.0005	0.0005	0.0004	0.0004	0.0004	0.0004	0.0004	0.0004	0.0003
-3.2	0.0007	0.0007	0.0006	0.0006	0.0006	0.0006	0.0006	0.0005	0.0005	0.0005
-3.1	0.0010	0.0009	0.0009	0.0009	0.0008	0.0008	0.0008	0.0008	0.0007	0.0007
-3.0	0.0013	0.0013	0.0013	0.0012	0.0012	0.0011	0.0011	0.0011	0.0010	0.0010
-2.9	0.0019	0.0018	0.0018	0.0017	0.0016	0.0016	0.0015	0.0015	0.0014	0.0014
-2.7	0.0035	0.0034	0.0033	0.0032	0.0031	0.0030	0.0029	0.0028	0.0027	0.0026
-2.6	0.0047	0.0045	0.0044	0.0043	0.0041	0.0040	0.0039	0.0038	0.0037	0.0036
-2.5	0.0062	0.0060	0.0059	0.0057	0.0055	0.0054	0.0052	0.0051	0.0049	0.0048
-2.4	0.0082	0.0080	0.0078	0.0075	0.0073	0.0071	0.0069	0.0068	0.0066	0.0064
-2.3	0.0107	0.0104	0.0102	0.0099	0.0096	0.0094	0.0091	0.0089	0.0087	0.0084
-2.2	0.0139	0.0136	0.0132	0.0129	0.0125	0.0122	0.0119	0.0116	0.0113	0.0110
-2.1	0.0179	0.0174	0.0170	0.0166	0.0162	0.0158	0.0154	0.0150	0.0146	0.0143
-2.0	0.0228	0.0222	0.0217	0.0212	0.0207	0.0202	0.0197	0.0192	0.0188	0.0183
-1.9	0.0287	0.0281	0.0274	0.0268	0.0262	0.0256	0.0250	0.0244	0.0239	0.0233
-1.8	0.0359	0.0351	0.0344	0.0336	0.0329	0.0322	0.0314	0.0307	0.0301	0.0294
-1.7	0.0446	0.0436	0.0427	0.0418	0.0409	0.0401	0.0392	0.0384	0.0375	0.0367
-1.6	0.0548	0.0537	0.0526	0.0516	0.0505	0.0495	0.0485	0.0475	0.0465	0.0455
-1.5	0.0668	0.0655	0.0643	0.0630	0.0618	0.0606	0.0594	0.0582	0.0571	0.0559
-1.4	0.0808	0.0793	0.0778	0.0764	0.0749	0.0735	0.0721	0.0708	0.0694	0.0681
-1.3	0.0968	0.0951	0.0934	0.0918	0.0901	0.0885	0.0869	0.0853	0.0838	0.0823
-1.2	0.1151	0.1131	0.1112	0.1093	0.1075	0.1056	0.1038	0.1020	0.1003	0.0985
-1.1	0.1357	0.1335	0.1314	0.1292	0.1271	0.1251	0.1230	0.1210	0.1190	0.1170
-1.0	0.1587	0.1562	0.1539	0.1515	0.1492	0.1469	0.1446	0.1423	0.1401	0.1379
-0.9	0.1841	0.1814	0.1788	0.1762	0.1736	0.1711	0.1685	0.1660	0.1635	0.1611
-0.8	0.2119	0.2090	0.2061	0.2033	0.2005	0.1977	0.1949	0.1922	0.1894	0.1867
-0.7	0.2420	0.2389	0.2358	0.2327	0.2296	0.2266	0.2236	0.2206	0.2177	0.2148
-0.6	0.2743	0.2709	0.2676	0.2643	0.2611	0.2578	0.2546	0.2514	0.2483	0.2451
-0.5	0.3085	0.3050	0.3015	0.2981	0.2946	0.2912	0.2877	0.2843	0.2810	0.2776
-0.4	0.3446	0.3409	0.3372	0.3336	0.3300	0.3264	0.3228	0.3192	0.3156	0.3121
-0.3	0.3821	0.3783	0.3745	0.3707	0.3669	0.3632	0.3594	0.3557	0.3520	0.3483
-0.2	0.4207	0.4168	0.4129	0.4090	0.4052	0.4013	0.3974	0.3936	0.3897	0.3859
-0.1	0.4602	0.4562	0.4522	0.4483	0.4443	0.4404	0.4364	0.4325	0.4286	0.4247
-0.0	0.5000	0.4960	0.4920	0.4880	0.4840	0.4801	0.4761	0.4721	0.4681	0.4641

Table of Standard Normal Probabilities for Positive Z-scores

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.991		
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.99		
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.99		
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.99		
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.99		
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.99		
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.99		
3.0	0.9987	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.99		
3.1	0.9990	0.9991	0.9991	0.9991	0.9992	0.9992	0.9992	0.99		
3.2	0.9993	0.9993	0.9994	0.9994	0.9994	0.9994	0.9994	0.99		
3.3	0.9995	0.9995	0.9995	0.9996	0.9996	0.9996	0.9996	0.99		
3.4	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.99		

Note that the probabilities given in this table represent the area to the LEFT of the z-score.
The area to the RIGHT of a z-score = 1 – the area to the LEFT of the z-score

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
-3.4	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0002
-3.3	0.0005	0.0005	0.0005	0.0004	0.0004	0.0004	0.0004	0.0004	0.0004	0.0003
-3.2	0.0007	0.0007	0.0006	0.0006	0.0006	0.0006	0.0006	0.0005	0.0005	0.0005
-3.1	0.0010	0.0009	0.0009	0.0009	0.0008	0.0008	0.0008	0.0008	0.0007	0.0007
-3.0	0.0013	0.0013	0.0013	0.0012	0.0012	0.0011	0.0011	0.0011	0.0010	0.0010
-2.9	0.0019	0.0018	0.0018	0.0017	0.0016	0.0016	0.0015	0.0015	0.0014	0.0014
-2.8	0.0026	0.0025	0.0024	0.0023	0.0023	0.0022	0.0021	0.0021	0.0020	0.0019
-2.7	0.0035	0.0034	0.0033	0.0032	0.0031	0.0030	0.0029	0.0028	0.0027	0.0026
-2.6	0.0047	0.0045	0.0044	0.0043	0.0041	0.0040	0.0039	0.0038	0.0037	0.0036

Re-scaling to standard normal to compare distributions

$$Z = \frac{x-\mu}{\sigma}$$

- x = distribution value of interest ("raw score")
- μ, σ = r.v./population mean, standard deviation

Example: Weight for a population of rabbits follows a normal distribution $N(2.6,\ 1.1)$

What is the Z-score for a 3 pound rabbit?

$$Z = \frac{x-\mu}{\sigma} = \frac{3-2.6}{\sqrt{1.1}} = \mathbf{0.381}$$

Does is make sense that this number is positive?

What is probability a rabbit weighs less than 3 pounds?

```
pnorm(0.381) = 0.648

pnorm(3, 2.6, sqrt(1.1)) = 0.648
```

THE FUTURE IS NOW

Normal distributions functions, revisited

All functions assume standard normal. Provide additional arguments for other normals:

Standard normal	Any normal
`pnorm(q) = pnorm(q, 0, 1)`	`pnorm(q, mean, sd)`

Z-scores are most useful for comparing different distributions

Weight for rabbit pop A is distributed $N(2.6,\ 1.1)$

Weight for rabbit pop B is distributed $N(2.9,\ 0.17)$

Which of these two rabbits is bigger? Pop A rabbit weighting 2.95 lbs, or pop B rabbit weighing 3.1 lbs?

Population A: $Z = \frac{x-\mu}{\sigma} = \frac{2.95-2.6}{\sqrt{1.1}} = 0.334$

Population B: $Z = \frac{x-\mu}{\sigma} = \frac{3.1-2.9}{\sqrt{0.17}} = 0.485$

Putting it all together

The height of European men is distributed as $N(175,\ \ 53.3)$

The height of European women is distributed as $N(162.5, 34.8)$

What proportion of men is shorter than 150 cm, aka P(man < 150)?

Using Z-scores

$$Z = \frac{x-\mu}{\sigma} = \frac{150-175}{\sqrt{53.3}} = \text{-}3.424$$

```
> pnorm(-3.424)
[1] 0.0003085331
```

Skipping Z-scores

```
> pnorm(150, 175, sqrt(53.3))
[1] 0.0003081516
```

Putting it all together

Men: $N(175, 53.3)$
Women: $N(162.5, 34.8)$

What proportion of women is taller than 162.5 cm? **50%**

Putting it all together

Men: $N(175, 53.3)$

Women: $N(162.5, 34.8)$

What proportion of women is taller than 170 cm?

Using Z-scores

$$Z = \frac{x-\mu}{\sigma} = \frac{170-162.5}{\sqrt{34.8}} = 1.2713$$

```
> 1 - pnorm(1.2713)
[1] 0.101811
```

Skipping Z-scores

```
> 1 - pnorm(170, 162.5,
sqrt(34.8))
[1] 0.1017987
```

Putting it all together

Men: $N(175, 53.3)$
Women: $N(162.5, 34.8)$

What is the tallest a woman can be and still be in the bottom 22%?

Using Z-scores

```
> qnorm(0.22)
[1]  -0.7721932
```

$Z = \frac{x-\mu}{\sigma}$ → $\mathrm{x} = Z\sigma + \mu$

$$= -0.7722 * \sqrt{34.8} + 162.5$$
$$= \boldsymbol{157.9\ cm}$$

Skipping Z-scores

```
> qnorm(0.22, 162.5, sqrt(34.8))
  [1] 157.9447
```

Putting it all together

Men: $N(175, 53.3)$
Women: $N(162.5, 34.8)$

What is the **shortest** a woman can be and still be in the **top** 22%?

Using Z-scores

```
> -1 * qnorm(0.22)
[1]  0.7721932
```

$$Z = \frac{x-\mu}{\sigma} \quad \rightarrow \quad \mathrm{x} = Z\sigma + \mu$$

$$= 0.7722 * \sqrt{34.8} + 162.5$$

$$= \mathbf{167.05\ cm}$$

Skipping Z-scores

```
> qnorm(1-0.22, 162.5, sqrt(34.8))
  [1] 167.0553
```

Putting it all together

Men: $N(175, 53.3)$
Women: $N(162.5, 34.8)$

What is the probability a randomly chosen man is between 175–182 cm tall?

→ P(X<182) – P(X<175) = P(X<182) – 0.5

```
> pnorm(182, 175, sqrt(53.3)) - 0.5
  [1] 0.3311738
```

Putting it all together

Men: $N(175, 53.3)$
Women: $N(162.5, 34.8)$

What is the probability a randomly chosen man is either between 175–182 cm tall or between 150—160 cm tall?

→ $P(175 < X < 182) + P(150 < X < 160)$

```
### First probability
> pnorm(182, 175, sqrt(53.3)) - 0.5
  [1] 0.3311738

> ### Second prob.
> pnorm(160, 175, sqrt(53.3)) - pnorm(150, 175, sqrt(53.3))
  [1] 0.01965059

> 0.3311738 + 0.01965059
  [1] 0.3508244
```

Putting it all together

Men: $N(175, 53.3)$
Women: $N(162.5, 34.8)$

I have two randomly-chosen European friends, one man and one woman each. What is the probability the man is at least 180 cm and the woman is between 163—170 cm?

→ P(man > 180) x P(163 < woman < 170)

```
### First probability
> 1 - pnorm(180, 175, sqrt(53.3))
  [1] 0.2467138

> ### Second prob.
> pnorm(170, 162.5, sqrt(34.8)) – pnorm(163, 162.5, sqrt(34.8))
  [1] 0.3644282

> 0.246713*0.3644282
  [1] 0.08990917
```

Putting it all together

Men: $N(175, 53.3)$
Women: $N(162.5, 34.8)$

I have two new randomly-chosen European friends, one man and one woman each. What is the probability the man is 180 cm and the woman is 163 cm?

→ P(man = 180) x P(woman = 163)

→ **0**

Putting it all together

Men: $N(175, 53.3)$
Women: $N(162.5, 34.8)$

Assume 50.8% of Europeans are women. If a randomly-chosen person is shorter than 155 cm tall, what is the probability the person is a woman?

→ P(woman | < 155) = P(<155 | woman) * P(woman) / P(<155)

0.102 0.508

```
### P(<155 | woman)
> pnorm(155, 162.5, sqrt(34.8))
  [1] 0.1017987
```

Solving the denominator

P(<155) = P(<155 and man) + P(<155 and woman) =

P(<155|man)*P(man) + P(<155|woman)*P(woman)

0.0031 0.492 0.102 0.508

= 0.0533

```
### P(<155 | man)
> pnorm(155, 175, sqrt(53.3))
  [1] 0.003076926
```

Putting it all together

Men: $N(175, 53.3)$
Women: $N(162.5, 34.8)$

Assume 50.8% of Europeans are women. If a randomly-chosen person is shorter than 155 cm, what is the probability the person is a woman?

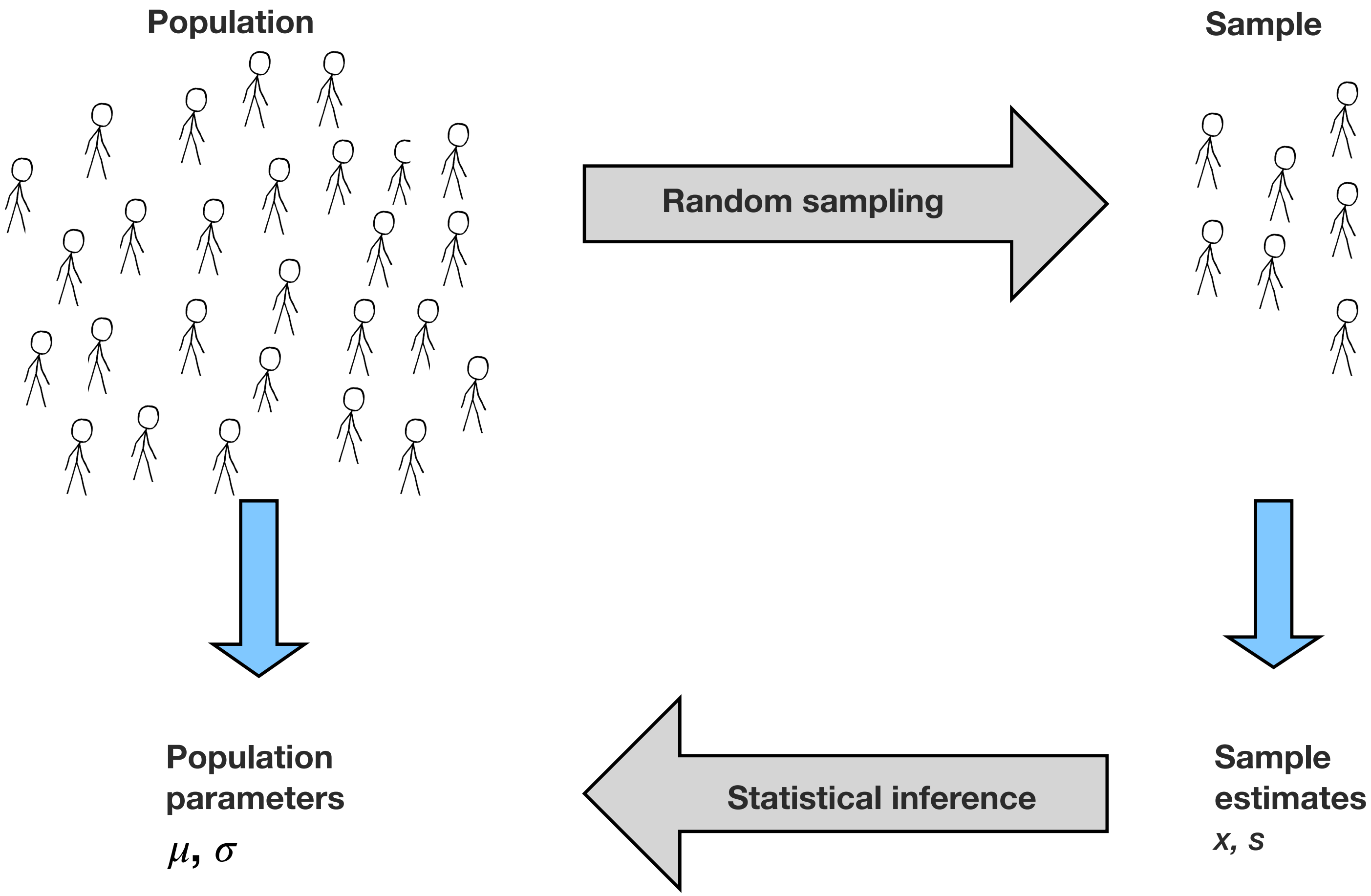
→ P(woman | < 155) = P(<155 | woman) * P(woman) / P(<155)

0.102 0.508 0.533

= 0.972

BREAK

Statistical inference



Population

Random sampling

Sample

Population parameters

μ, σ

Statistical inference

Sample estimates

x, *s*

Two main flavors of statistical inference

Estimation

- Estimate a population parameter from sample data
- Point estimates: What is the population mean?
- Interval estimates: In what range of values is the population mean likely to fall?

Hypothesis testing

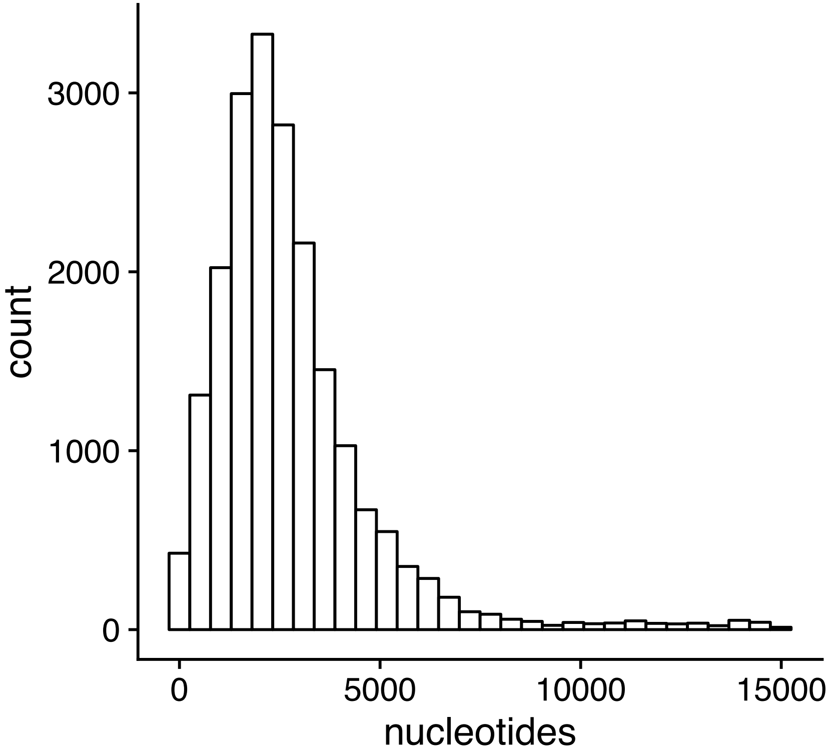
- Test whether the value of a population parameter is equal to some specific value
- Is there evidence that my sample differs from some underlying population?

The sampling distribution

The probability distribution of values for an estimate that we obtain under sampling

Obtaining a sampling distribution

```
> genes <- read.csv("genes.csv")
> head(genes)
  nucleotides
1        3785
2        7416
3        2135
4        7682
5        5766
6       11079

> mean(genes$nucleotides)
[1] 2761.039
> sd(genes$nucleotides)
[1] 2037.645
```

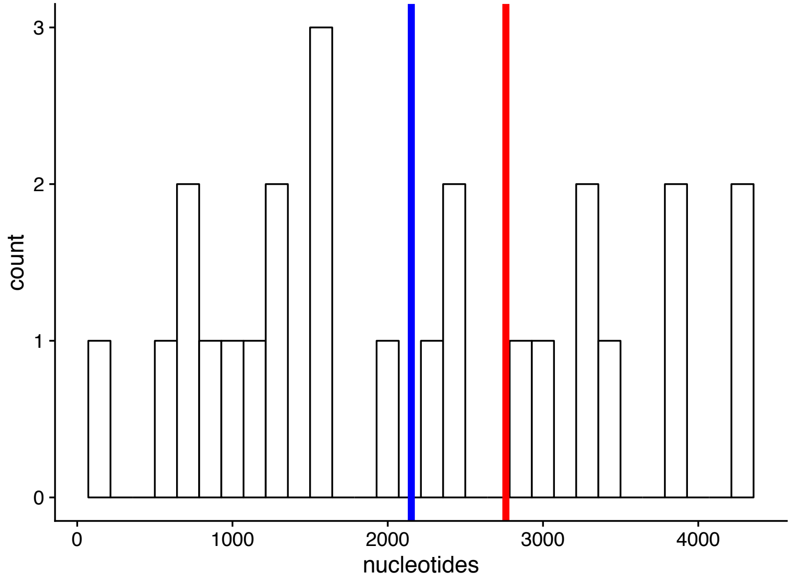


```
> ggplot(genes, aes(x=nucleotides)) +
geom_histogram(fill="white", color="black")
```

Obtaining a sampling distribution

```
### the function sample_n draws a random sample of rows

> small.sample <- genes %>% sample_n(25)
> mean(small.sample$nucleotides)
   [1] 2151.8

> ggplot(small.sample , aes(x = nucleotides)) +
       geom_histogram() +
       geom_vline(xintercept=2151.8, color="blue") +
       geom_vline(xintercept= 2761.039, color="red")
```

The sample mean for a random sample of N=25 is $\bar{x} = 2151.8$



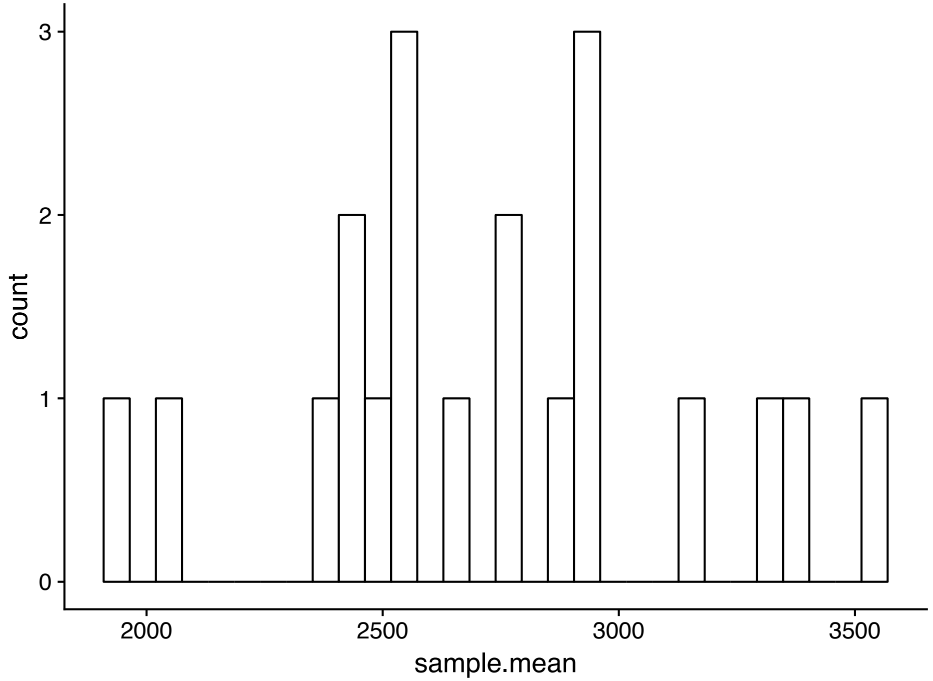
```
geom_vline(xintercept=…)
geom_hline(yintercept=…)
geom_abline(yintercept=…, slope=…)
```

Obtaining a sampling distribution

Now imagine we draw **20** samples of N=25 and compute each of their means:

```
> head(n20.means)
  sample.mean
1     2584.84
2     2574.12
3     2382.64
4     3143.68
5     2252.56
6     2368.44
```

Sampling distribution of the mean



Quantifying the sampling distribution

The standard error is the standard deviation of the estimate of the sampling distribution

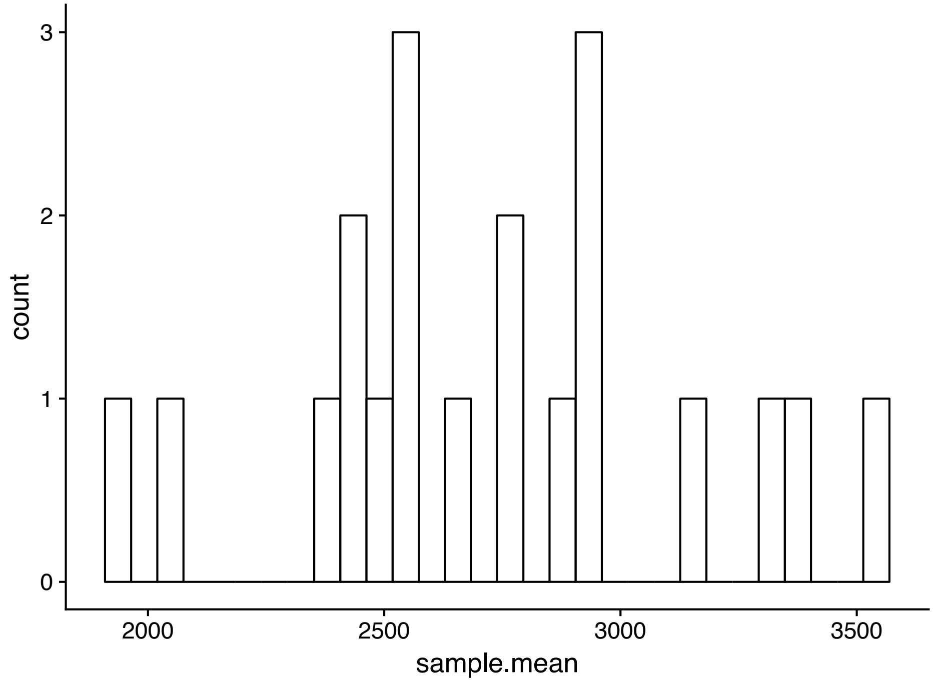
- Standard error of the mean: $SE_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$, approximate with $\frac{s}{\sqrt{n}}$
- SE is **not** the standard deviation of a sample
- Here, n represents the <u>number of samples</u> (**not** the sample size)

It also quantifies the **precision of our estimate**, i.e. how far from the population parameter we are

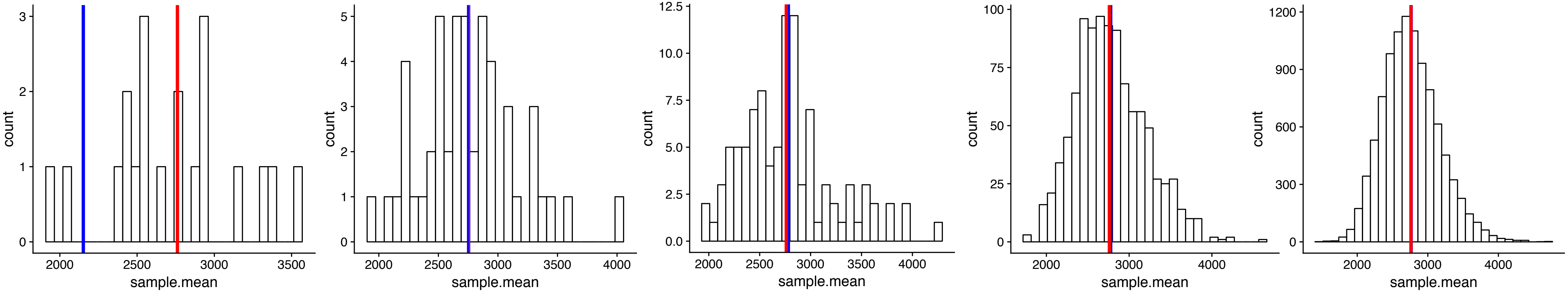
Computing the standard error of the mean

```
> head(n20.means)
  sample.mean
1     2584.84
2     2574.12
3     2382.64
4     3143.68
5     2252.56
6     2368.44

> sd(n20.means$sample.mean) / sqrt(20)
 [1] 93.11888
```

Sampling distribution of the mean



Several sampling distributions comprised of N samples, each of n=25



N=20

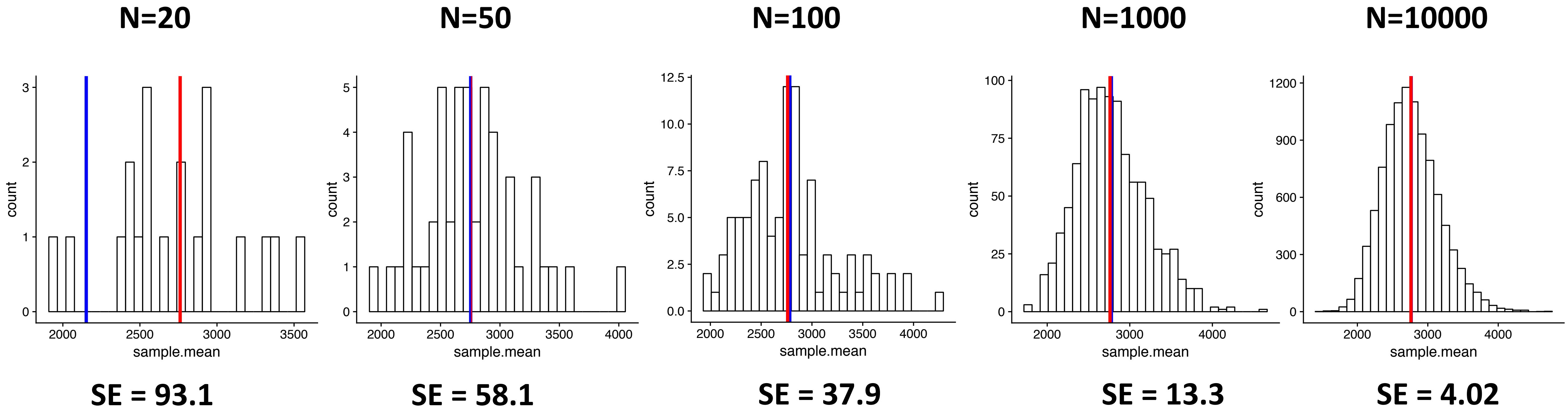
N=50

N=100

N=1000

N=10000

Standard error decreases as N increases



N=20

SE = 93.1

N=50

SE = 58.1

N=100

SE = 37.9

N=1000

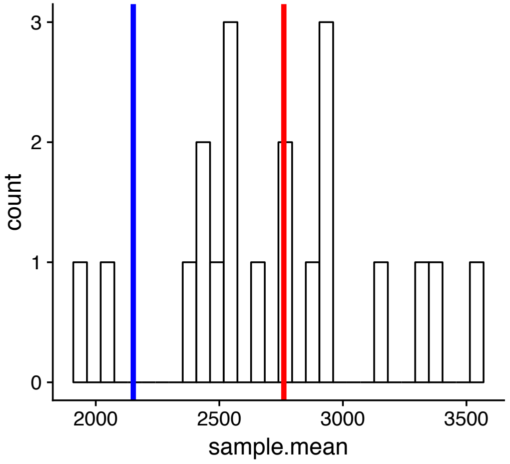
SE = 13.3

N=10000

SE = 4.02

Therefore, mean of sampling distribution approaches population mean 2761.039

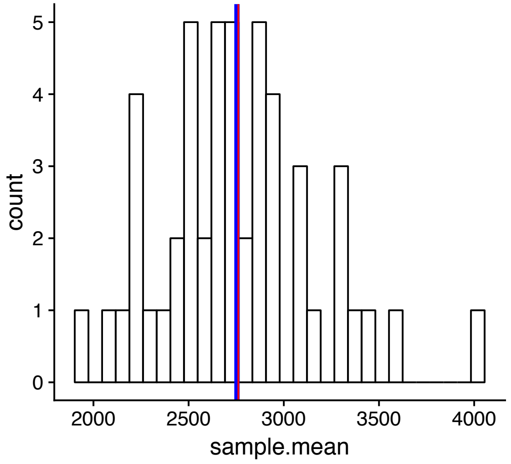
N=20



SE = 93.1

$\overline{x}$ = 2780.89

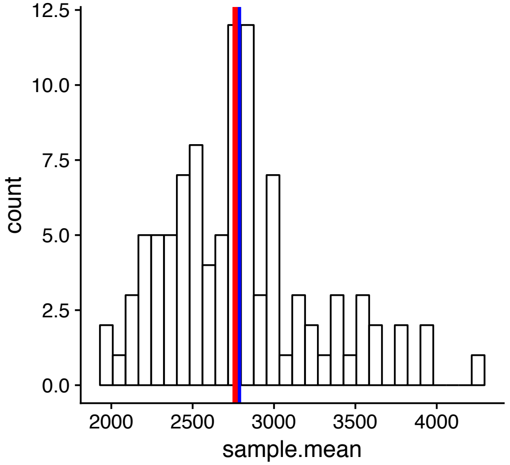
N=50



SE = 58.1

$\overline{x}$ = 2753.91

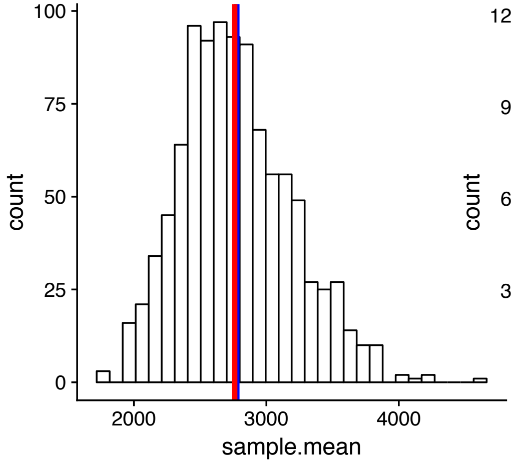
N=100



SE = 37.9

$\overline{x}$ = 2781.51

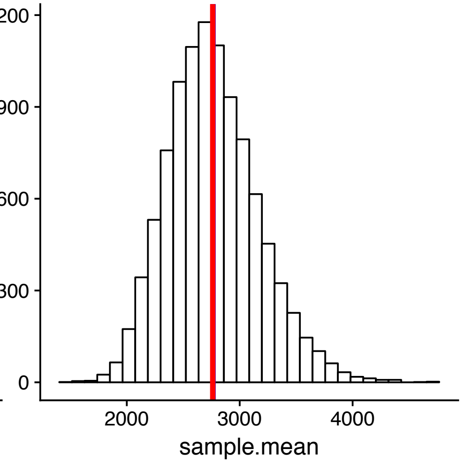
N=1000



SE = 13.3

$\overline{x}$ = 2777.02

N=10000

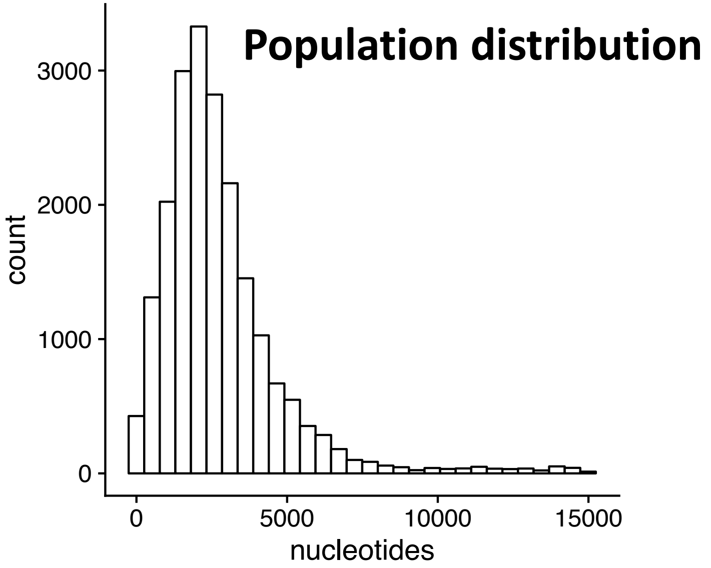


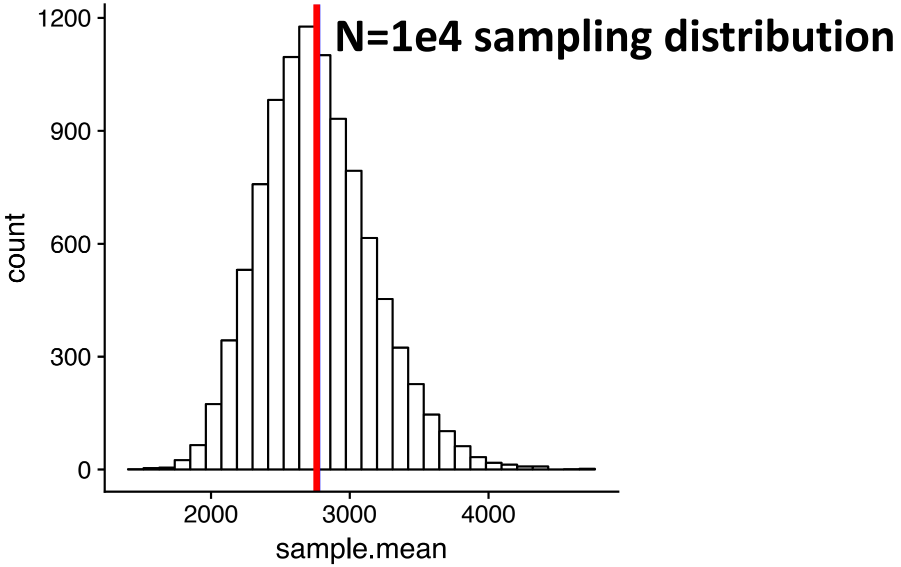
SE = 4.02

$\overline{x}$ = 2763.82

The Central Limit Theorem

As sample size increases, **the sampling distribution of the mean** will be approximately **normal** regardless of true population distribution





Next week..

Introduction to hypothesis testing and comparing means

More fun facts on estimation will come later in the semester, to be bundled with *likelihood*