STAT 311: LECTURE 14

Heavily based on lecture notes from Martina Morris

Continuous Random Variables

Logistics

- Sam's office hours will be moved to Friday 1-3

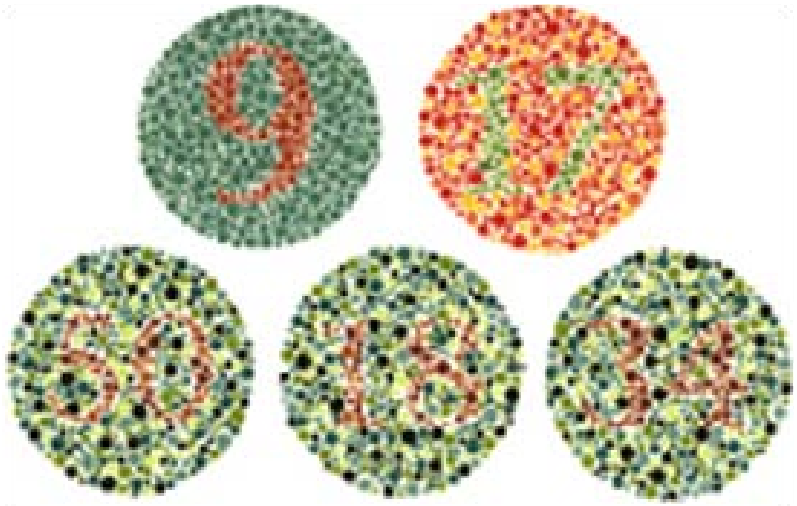
Discrete random variables

- Examples we looked at included
  - Bernoulli (0,1) outcomes
  - Binomial – counts, sums of independent Bernoulli RVs
  - Poisson – counts, when expressed as a rate (per time, per capita, per class …)

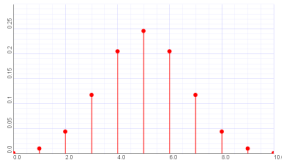
- We have formal expressions for the PDFs/CDFs
  - How do we use these?

Example: Color Blindness

The frequency of color blindness (dyschromatopsia) in the Caucasian American male population is estimated to be about 8%.



Probabilities for specific values



- Say you have 20 men from this population.
- What is the probability that **exactly 5** will be color blind?

$$P(X = 5) = b(5; 20, .08) = \binom{20}{5} 0.08^5 (1 - 0.08)^{15}$$

$$= 0.015$$

So, only a 1.5% chance

- How to calculate or find this?
  - Calculate directly, as above
  - Software: R: dbinom(k, n, p)

Note what we have just done

- We can now tell you the probability that any discrete random variable you choose takes on a certain value
- And what value we expect it to take
- And what fraction of the time it will take a value more or less than this
- And more … (a lot more, by the end of the course)

You just tell us about the process, the possible outcomes, and the value of the governing parameter(s) (if any)

7

Continuous random variables

Uniform, Exponential and Normal distributions

What changes

- **The sample spaces can not be enumerated**
  - There are an infinite number of possible outcomes
  - And the probability of any single outcome is effectively 0

- As a result: All of our probability calculations are **restricted to cumulative probabilities**
  - $P(X < x)$ or
  - $P(X > x)$ or
  - $P(X \leq x)$

The PDF

- **The PDF for continuous variables is not the probability of each value**
  - That probability = 0
- It is the "probability density"
  - Defined by the law of total probability: $\int_a^b f(x)dx = 1$
  - Where a and b are the range of the RV: $x \in [a, b]$
- So, be warned
  - You will see that $f(x)$ can take values > 1 in some contexts…

Intuition behind pdf of continuous variables

- What is the probability someone is 1.5 meters tall?
  - What you really mean is what is the probability someone is taller than 1.45 meters and shorter than 1.55 meters tall
- Let's get more more precise, what is the probability someone is 1.50 meters tall?
  - What you really mean is what is the probability someone is taller than 1.495 meters and shorter than 1.505 meters tall

Intuition behind pdf of continuous variables

- Since continuous variables should be measured to arbitrary precision, let's be very precise. What is the probability someone is

  1.500000000000000000000000000

  meters tall?
  - This is starting to get pretty small
- Imagine letting the number of 0's grow
- We start to think about only a single point, and the probability goes to 0

Expected Values and Variances

The definitions use integrals (instead of sums):

$$E(X) = \mu_x = \int_a^b x f(x) dx$$

$$Var(X) = \sigma_x^2 = \int_a^b (x - \mu)^2 f(x) dx$$

We will not be deriving these the way we did for the Bernoulli and Binomial, but you should know that the derivations depend on integrals.

What doesn't change

- There is an underlying stochastic process
  - Here we're picking a random continuous value
  - Independence still matters

- That process defines
  - The sample space (the range of possible outcomes)
  - The number of outcomes (infinite, given the continuous value)
  - The distribution of outcomes
    - That we can represent with a mathematical function
    - And derive expected values, variances, PDFs and CDFs

Simplest case: Uniform distributions

Example from UH: Waiting times in an interval

- The bus comes exactly every 10 minutes
- People arrive at the stop at random times
  - Independently
- What is the distribution of their waiting times?

What are the key probability elements?

- What is the sample space?
  - A range: [0 – 10] minutes
- How many possible outcomes?
  - An infinite number of outcomes
- How are outcomes distributed across the range?
  - What does “at random” imply for the probability of each value?
  - If you repeatedly sampled a number at random from the [0 – 10] range
    - What would the distribution of sample values look like?
  - What’s your best guess of the waiting time?

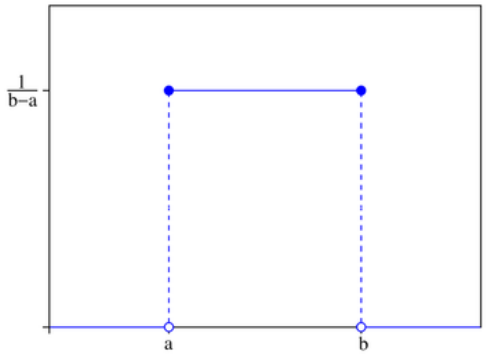
The Uniform distribution

- Every outcome has the same probability

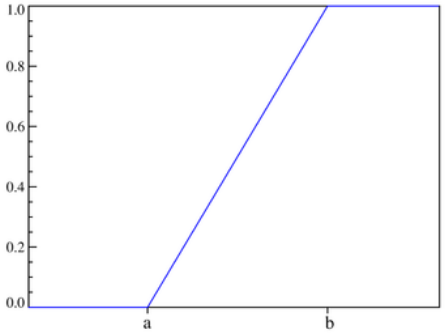
X ~ Uniform(a, b)

$$\text{If } a \le x \le b$$
$$f(x) = \frac{1}{b-a}$$
$$\text{else}$$
$$f(x) = 0$$

$$\mu_x = \frac{a+b}{2} \qquad \sigma_x^2 = \frac{(b-a)^2}{12}$$



PDF



CDF

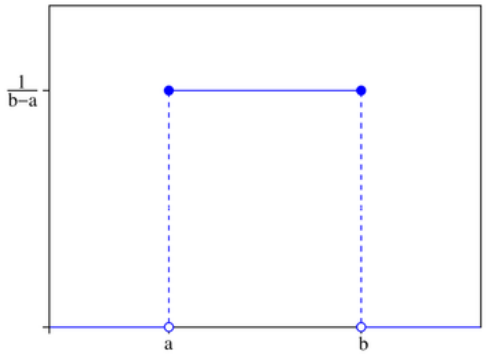
$f(x)$ and the Uniform distribution

- What if a=0 and b=1/2?

$$f(x;\, a,b) = \frac{1}{(b - a)}$$

$$= \frac{1}{(0.5 - 0)}$$

$$= 2$$



PDF

Good example for remembering that $f(x)$ is not a probability for continuous RVs, it's the probability density

Next: a process defined by a rate

- Imagine a different stochastic process
  - Now, there is no defined interval of time
  - Instead, the process is defined by a rate at which events happen

- I receive emails at a rate of about 2 / hr

- What is the distribution of waiting times until my next email

What are the key probability elements?

- **What is the sample space?**
  - All positive values: $[0, \infty)$ minutes
- **How many possible outcomes?**
  - An infinite number of outcomes
- **How are outcomes distributed across the range?**
  - Do all outcomes have the same probability now?
  - Given the rate of email arrivals (2 / hr), what is your best guess of your wait time when you will receive you next email?
  - If you've waited 1 hr, what is you best guess of how much longer you'll wait?

Why doesn't your expected wait time change?

- It's a special property of the exponential distribution
  - Because the right tail of this distribution is unbounded
  - And the distribution is "memoryless"

- Just like a run of heads in a sequence of coin tosses
  - Once you condition on the previous sequence, the probability of a head on the next toss is just 0.5, same as every other toss.

- Here, once you condition on having waited as long as you already have
  - The mean wait time in the rest of the tail is just what it was before

The Exponential distribution

- The waiting time distribution for events that have a Poisson rate distribution

$$X \sim \text{Exponential}(\lambda)$$

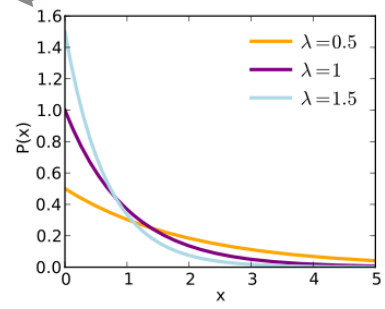
If x > 0

$$f(x;\lambda) = \lambda e^{-\lambda x}$$

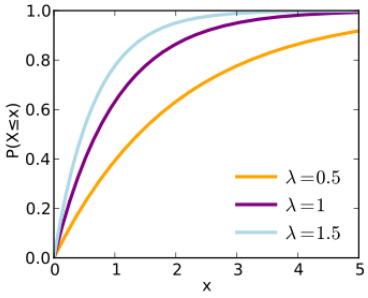
$$\mu_x = \frac{1}{\lambda}$$

$$\sigma_x^2 = \frac{1}{\lambda^2}$$

Note: $f(x) > 1$!



PDF



CDF

The Poisson and the Exponential

- Poisson: events per time unit
- Exponential: time units per event
- They share the same rate parameter: λ

Poisson: expected number of emails per hour = 2 = 1 every 30 minutes

Exponential: expected waiting time to the next email = $\frac{1}{2}$ hour

$$\mu_{\text{poisson}} = \frac{1}{\mu_{\text{exponential}}}$$

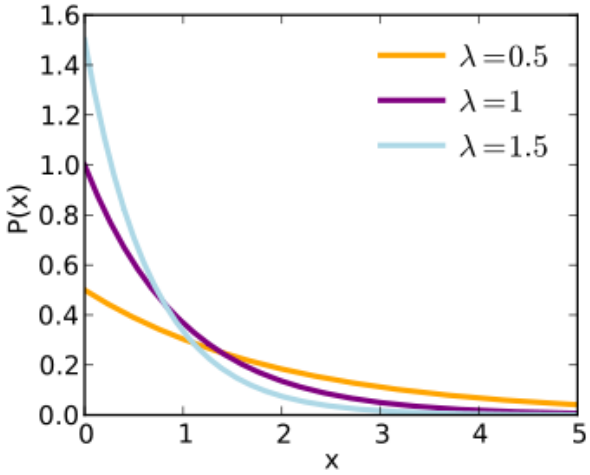
Wait times are "memoryless" for others too

- When John gets to the bus stop, his expected wait is $\frac{1}{\lambda}$
- If I get to the stop after him, my expected wait is also $\frac{1}{\lambda}$
- No matter when you get to the bus stop, your expected wait is $\frac{1}{\lambda}$

Because the events are independent

Weird, but

- Note how the PDF changes as a function of λ
- As the rate of events increases
- The PDF gets steeper
- Why?
  - Because busses can get there very late
  - But getting there earlier is bounded by 0
  - So the waiting times heap up at the lower boundary



As λ decreases, what does the PDF begin to remind you of?

As decreases, what does the PDF begin to remind you of?

Finally: The Normal distribution

- Is there a stochastic process that defines the Normal distribution?

- Yes, but
  - The derivation is in your supplemental readings (it's an optional assignment)
  - … Let's just say it is neither simple nor intuitive

- Intuition: it is the result you get from combining many different processes (e.g., a sum of many variables)
  - This gives rise to a symmetric bell-shaped curve on a continuous scale

Examples of approximately Normal RVs

- Distributions of:
  - Height and weight in the population
    - A function of many different underlying processes: genetics, nutrition, environment
  - Monthly average rainfall
    - A function of many different climactic processes: temperature, evaporation, atmospheric pressure

- But not:
  - Distribution of income in the US
    - (too skewed)
  - Distribution of years of educational attainment
    - (integer valued, a limited range, non-negative)

Combinations of RVs we have seen

Linear combinations of independent RVs:

- Roll a die 20 times, and take the sum of the face values you get
  - Do this 100 times, and look at the distribution of the sums
- Count the average number of busses per hour on a given day
  - Do this for a month, and look at the distribution of the means
- Calculate the average waiting time to your next email today
  - Look at the distribution of the averages across the last month

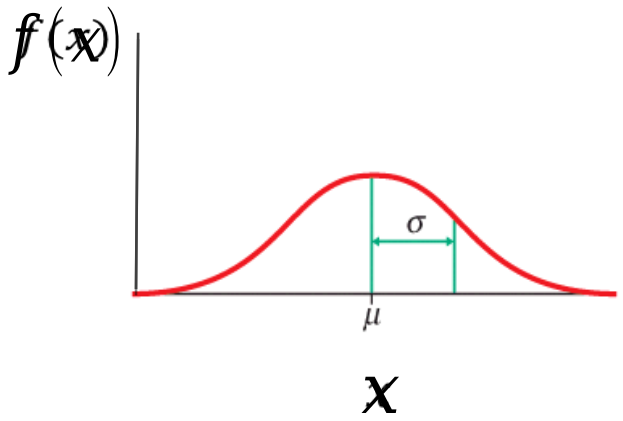
Warning:

- Some combinations approach normality
- Some do not
- So you can't just assume all continuous distributions, or all combinations of variables have a Normal distribution

The Normal distribution

- One of the most commonly used distributions in statistics.
- Maps x to the probability density of x

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

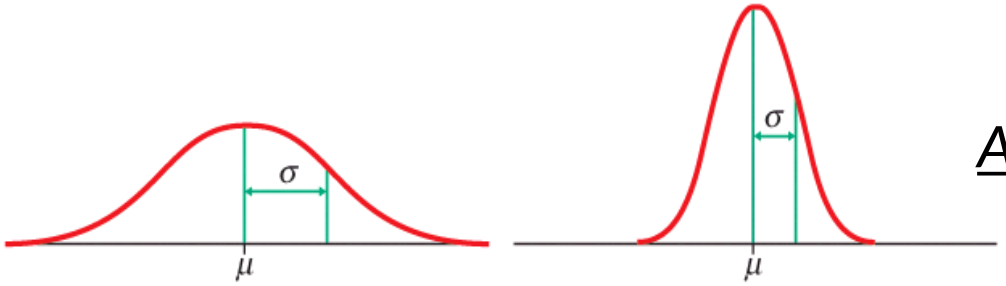


- What elements of this function do you recognize now?
  - So, this is saying the probability of a value is a function of its zscore...

Elements of this function

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

$$X \sim N(\mu_X, \sigma_X)$$



Because this is a probability density curve, the area under the curve sums to 1 (or 100%)

The Random Variable: **X**

***2 constants:* e and π**

e = 2.7182… The base of the natural logarithm

π = pi = 3.1415…

And two free parameters:

μ = the mean of X

σ = the standard deviation of X

The two free parameters: μ and σ

- The entire PDF is specified by these
  - While it is a complex mapping of x to $f(x)$
  - These two parameters are all you need to know to calculate the probability of every range of values.

- And because the Normal distribution is symmetric
  - σ tells us a great deal about the cumulative probability of the data.
  - We express this in terms of the fraction of the data within 1, 2 or 3 standard deviations from the mean

There are 2 special properties of Normals

- The “empirical rule”
  - Tells us the approximate fraction of values in X that lie within 1, 2 or 3 standard deviations of the mean

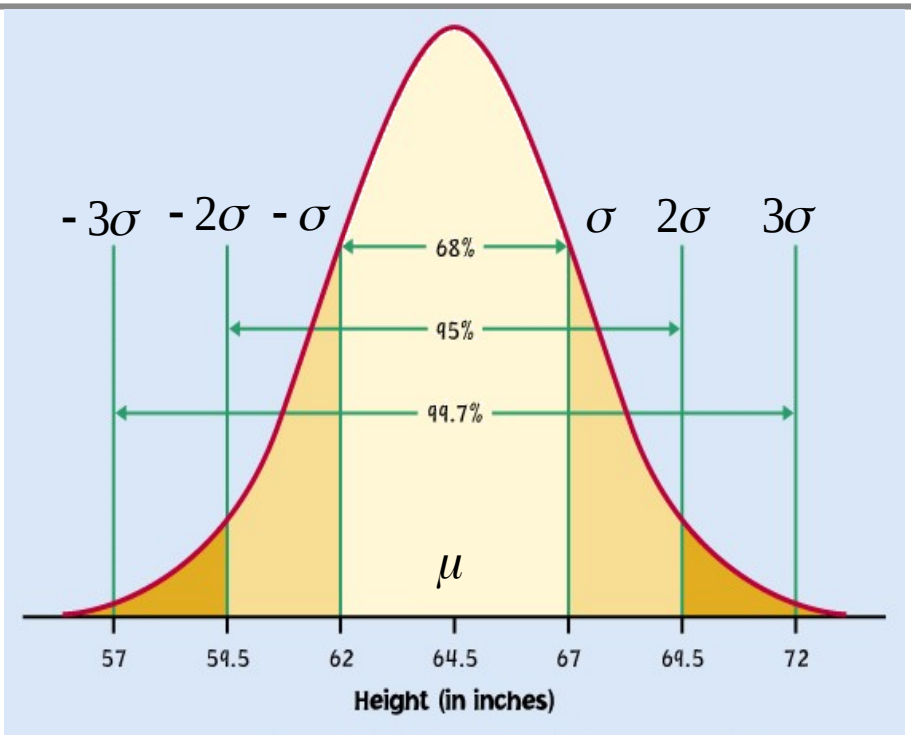
- All Normal distributions can be transformed into the “standard Normal”
  - The standard Normal has mean=0, and standard deviation=1

1. The empirical rule for Normal Distributions

Approximately:

- **68%** of all observations are within ±1σ of the mean (μ).
- **95%** are within ±2σ of μ.
- **99.7%** are within ±3σ of μ.

The remaining 0.3% is split between the upper and lower tails (half in each)

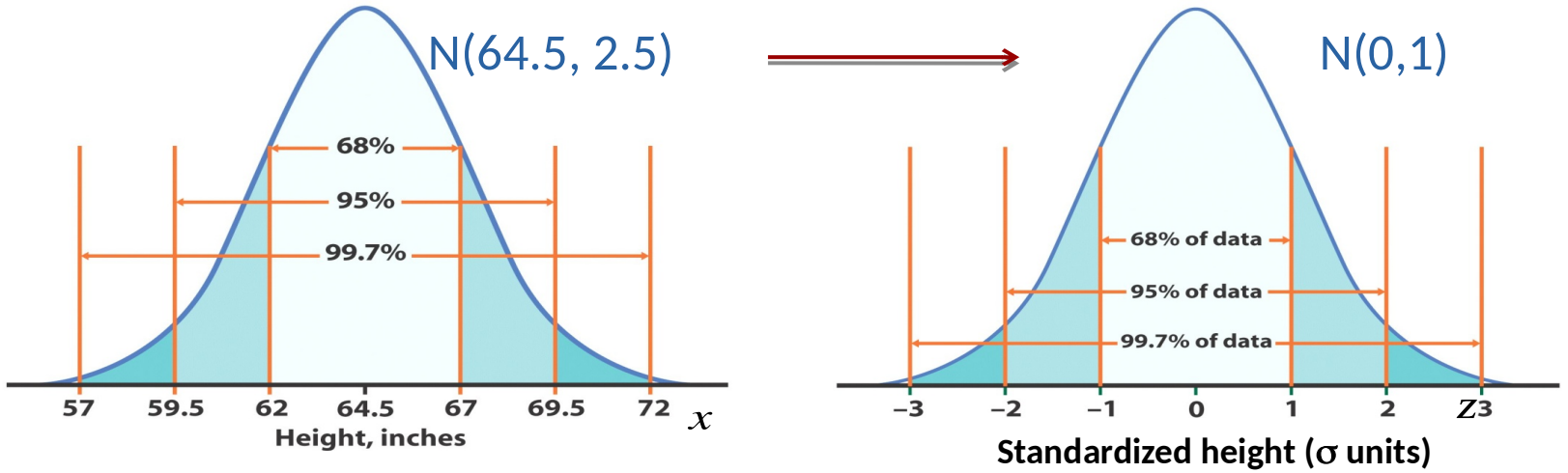


mean μ = 64.5 standard deviation σ = 2.5

N(μ, σ) = N(64.5, 2.5)

2. The standard Normal distribution

Because all Normal distributions share these properties, we can transform any Normal distribution N(μ,σ) into the *"standard Normal distribution"* N(0,1).



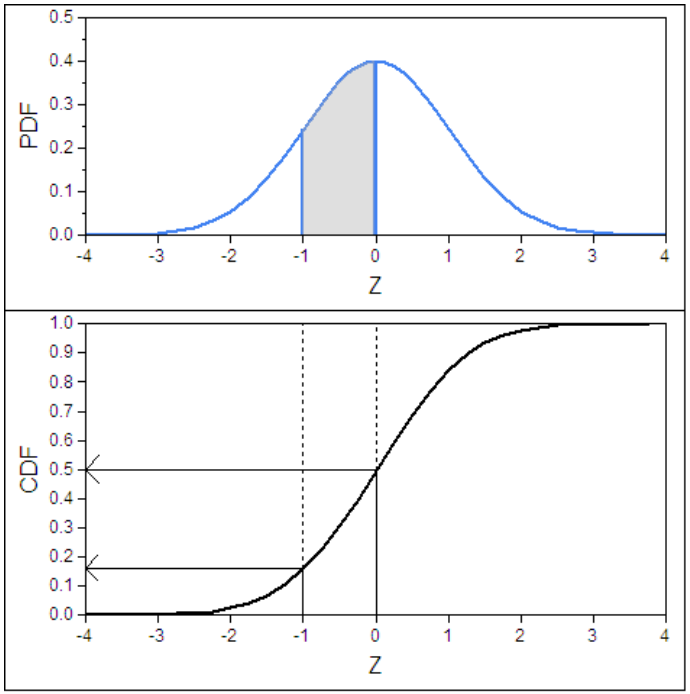
We do this by transforming the data values into their z-scores: $z = \frac{x - \mu}{\sigma}$

Calculating probabilities

For continuous RVs

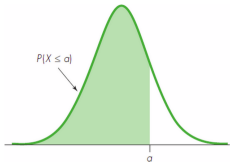
The basic idea is the same

- You can use software
- Or a Table lookup
- Or in some special cases, like the Normal, you can use the empirical rule
- But remember, we only calculate cumulative probabilities for continuous variables



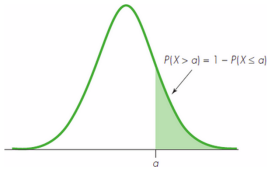
For cumulative probabilities

- There are three basic measures for $P(X)$



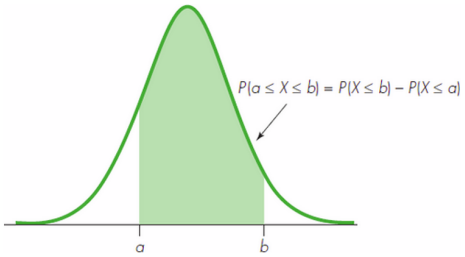
$P(X \leq a)$

The probability that X is less or equal to a certain value



$P(X > a)$

The probability that X is greater than a certain value



$P(a \leq X \leq b)$

The probability that X lies in the range $[a, b]$

Table lookups

- **The most common lookup is for a Normal distribution**
  - **And for this, we always use a standard Normal table**
- **So you need to transform your value into a zscore to look it up**
- Example: you want to lookup $P(X > a)$ for $X \sim N(\mu, \sigma)$
  - Calculate the zscore for the value: $z(a) = \frac{a-\mu}{\sigma}$
  - **And lookup the zscore in the table**

Using the standard Normal table

Table A gives the area under the standard Normal curve to the left of any z value.

zscore, 1st two digits

zscore, last digit

TABLE A Standard normal probabilities

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
−3.4	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0002
−3.3	.0005	.0005	.0005	.0004	.0004	.0004	.0004	.0004	.0004	.0003
−3.2	.0007	.0007	.0006	.0006	.0006	.0006	.0006	.0005	.0005	.0005
−3.1	.0010	.0009	.0009	.0009	.0008	.0008	.0008	.0008	.0007	.0007
−3.0	.0013	.0013	.0013	.0012	.0012	.0011	.0011	.0011	.0010	.0010
−2.9	.0019	.0018	.0018	.0017	.0016	.0016	.0015	.0015	.0014	.0014
−2.8	.0026	.0025	.0024	.0023	.0023	.0022	.0021	.0021	.0020	.0019
−2.7	.0035	.0034	.0033	.0032	.0031	.0030	.0029	.0028	.0027	.0026
−2.6	.0047	.0045	.0044	.0043	.0041	.0040	.0039	.0038	.0037	.0036
−2.5	.0062	.0060	.0059	.0057	.0055	.0054	.0052	.0051	.0049	.0048
−2.4	.0082	.0080	.0078	.0075	.0073	.0071	.0069	.0068	.0066	.0064
−2.3	.0107	.0104	.0102	.0099	.0096	.0094	.0091	.0089	.0087	.0084
−2.2	.0139	.0136	.0132	.0129	.0125	.0122	.0119	.0116	.0113	.0110
−2.1	.0179	.0174	.0170	.0166	.0162	.0158	.0154	.0150	.0146	.0143
−2.0	.0228	.0222	.0217	.0212	.0207	.0202	.0197	.0192	.0188	.0183
−1.9	.0287	.0281	.0274	.0268	.0262	.0256	.0250	.0244	.0239	.0233
−1.8	.0359	.0351	.0344	.0336	.0329	.0322	.0314	.0307	.0301	.0294
−1.7	.0446	.0436	.0427	.0418	.0409	.0401	.0392	.0384	.0375	.0367
−1.6	.0548	.0537	.0526	.0516	.0505	.0495	.0485	.0475	.0465	.0455
−1.5	.0668	.0655	.0643	.0630	.0618	.0606	.0594	.0582	.0571	.0559
−1.4	.0808	.0793	.0778	.0764	.0749	.0735	.0721	.0708	.0694	.0681

Cumulative probability that $Z \leq z$

Using the standard Normal table

Table A gives the area under the standard Normal curve to the left of any z value.

If your zscore = -2.40

0.82% of the N(0,1) density lies to the left of **z = -2.40**

TABLE A Standard normal probabilities

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
−3.4	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0002
−3.3	.0005	.0005	.0005	.0004	.0004	.0004	.0004	.0004	.0004	.0003
−3.2	.0007	.0007	.0006	.0006	.0006	.0006	.0006	.0005	.0005	.0005
−3.1	.0010	.0009	.0009	.0009	.0008	.0008	.0008	.0008	.0007	.0007
−3.0	.0013	.0013	.0013	.0012	.0012	.0011	.0011	.0011	.0010	.0010
−2.9	.0019	.0018	.0018	.0017	.0016	.0016	.0015	.0015	.0014	.0014
−2.8	.0026	.0025	.0024	.0023	.0023	.0022	.0021	.0021	.0020	.0019
−2.7	.0035	.0034	.0033	.0032	.0031	.0030	.0029	.0028	.0027	.0026
	.0047	.0045	.0044	.0043	.0041	.0040	.0039	.0038	.0037	.0036
−2.5	.0062	.0060	.0059	.0057	.0055	.0054	.0052	.0051	.0049	.0048
−2.4	.0082	.0080	.0078	.0075	.0073	.0071	.0069	.0068	.0066	.0064
−2.3	.0107	.0104	.0102	.0099	.0096	.0094	.0091	.0089	.0087	.0084
−2.2	.0139	.0136	.0132	.0129	.0125	.0122	.0119	.0116	.0113	.0110
−2.1	.0179	.0174	.0170	.0166	.0162	.0158	.0154	.0150	.0146	.0143
−2.0	.0228	.0222	.0217	.0212	.0207	.0202	.0197	.0192	.0188	.0183
−1.9	.0287	.0281	.0274	.0268	.0262	.0256	.0250	.0244	.0239	.0233
−1.8	.0359	.0351	.0344	.0336	.0329	.0322				
−1.7	.0446	.0436	.0427	.0418	.0409	.0401				
−1.6			.0526	.0516	.0505	.0495	.0485	.0475	.0465	.0455
−1.5			.0643	.0630	.0618	.0606	.0594	.0582	.0571	.0559
−1.4	.0808	.0793	.0778	.0764	.0749	.0735	.0721	.0708	.0694	.0681

0.80% < **-2.41**

0.69% < **-2.46**

Note

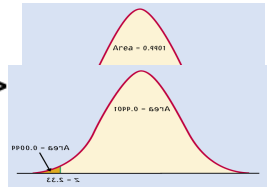
- **Many standard Normal tables only show the lower tail probabilities**

- **So if you want values from the upper tail, remember**

$$P(Z \leq -z) \text{ for } z \leq 0$$

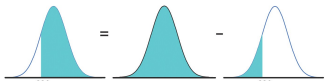
  - You can use the symmetry of the normal for values of $z >$

$$P(Z > z) = P(Z \leq -z) \text{ for } z > 0$$

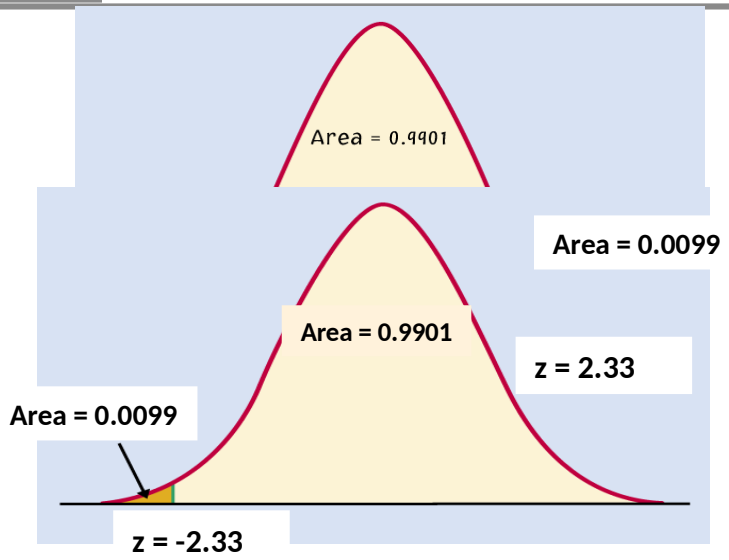


  - Or the complement rule for values of $z < 0$

$$P(Z > z) = 1 - P(Z \leq -z) \text{ for } z < 0$$



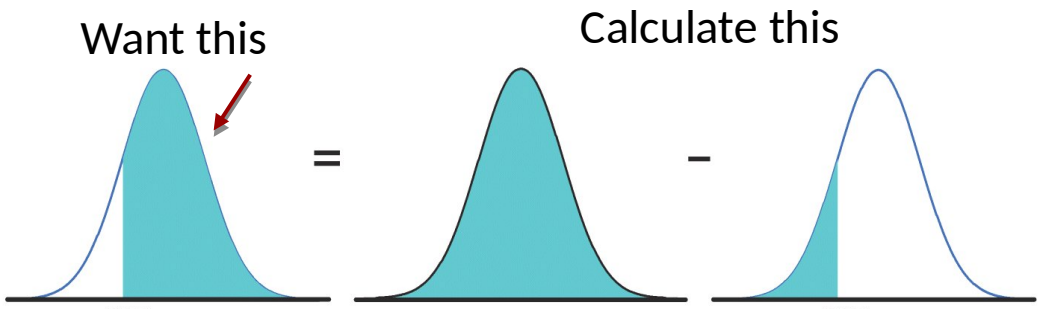
Looking up $P(Z > z)$ in a partial table



For z-scores above the mean:

$$P(Z > z) = P(Z < -z)$$

Because the Normal distribution is symmetric

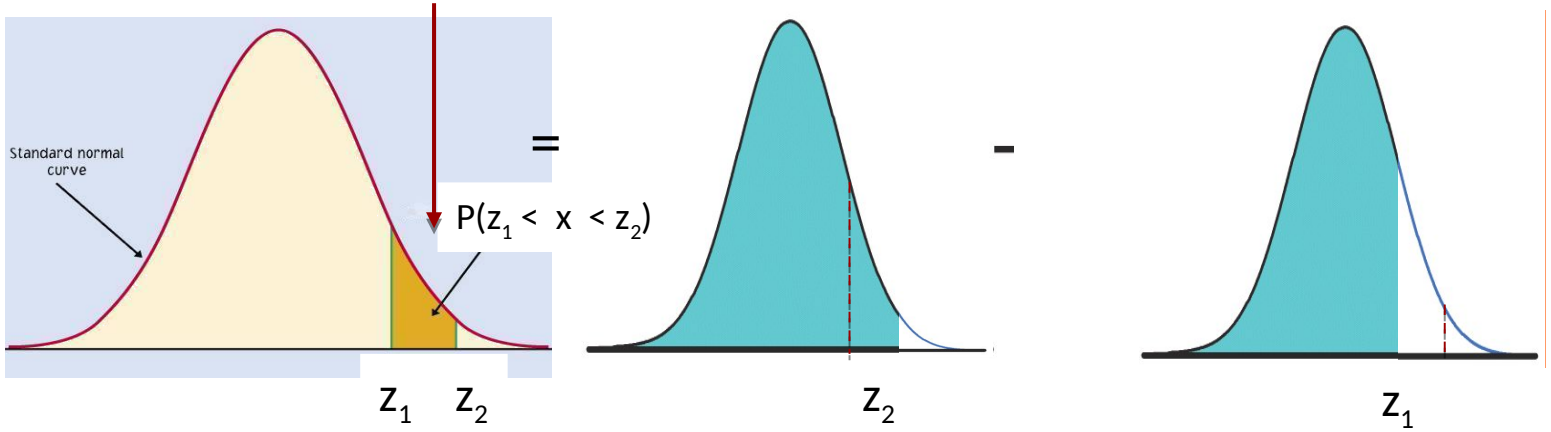


For z-scores below the mean:

$$P(Z > z) = 1 - P(Z < -z)$$

Using the complement rule

Calculating for the upper tail



$$P(z_1 < Z < z_2) = P(Z < z_2) - P(Z < z_1)$$

Then use symmetry and the complement rules for these

One more thing we use this for

- To identify the quantiles of a distribution
- Instead of asking:

"What is the probability that $X \leq a$?"

"What is the probability that ?"

- We can ask:

"What value of X defines the lower 5^{th} percentile of the distribution"

"What value of X defines the lower 5^{th} percentile of the distribution"

- This is just a reverse lookup
  - Just remember that zscores are quantiles

zscore = Inverse Normal CDF = quantile

- What zscore value defines the 5th percentile?
- Lookup 0.05 in the center of the table
- And map it to the zscore
- Here, z = -1.645 (using extrapolation)

TABLE A Standard normal probabilities

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
−3.4	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0002
−3.3	.0005	.0005	.0005	.0004	.0004	.0004	.0004	.0004	.0004	.0003
−3.2	.0007	.0007	.0006	.0006	.0006	.0006	.0006	.0005	.0005	.0005
−3.1	.0010	.0009	.0009	.0009	.0008	.0008	.0008	.0008	.0007	.0007
−3.0	.0013	.0013	.0013	.0012	.0012	.0011	.0011	.0011	.0010	.0010
−2.9	.0019	.0018	.0018	.0017	.0016	.0016	.0015	.0015	.0014	.0014
−2.8	.0026	.0025	.0024	.0023	.0023	.0022	.0021	.0021	.0020	.0019
−2.7	.0035	.0034	.0033	.0032	.0031	.0030	.0029	.0028	.0027	.0026
−2.6	.0047	.0045	.0044	.0043	.0041	.0040	.0039	.0038	.0037	.0036
−2.5	.0062	.0060	.0059	.0057	.0055	.0054	.0052	.0051	.0049	.0048
−2.4	.0082	.0080	.0078	.0075	.0073	.0071	.0069	.0068	.0066	.0064
−2.3	.0107	.0104	.0102	.0099	.0096	.0094	.0091	.0089	.0087	.0084
−2.2	.0139	.0136	.0132	.0129	.0125	.0122	.0119	.0116	.0113	.0110
−2.1	.0179	.0174	.0170	.0166	.0162	.0158	.0154	.0150	.0146	.0143
−2.0	.0228	.0222	.0217	.0212	.0207	.0202	.0197	.0192	.0188	.0183
−1.9	.0287	.0281	.0274	.0268	.0262	.0256	.0250	.0244	.0239	.0233
−1.8	.0359	.0351	.0344	.0336	.0329	.0322	.0314	.0307	.0301	.0294
−1.7	.0446	.0436	.0427	.0418	.0409	.0401	.0392	.0384	.0375	.0367
−1.6	.0548	.0537	.0526	.0516	.0505	.0495	.0485	.0475	.0465	.0455
−1.5	.0668	.0655	.0643	.0630	.0618	.0606	.0594	.0582	.0571	.0559
−1.4	.0808	.0793	.0778	.0764	.0749	.0735	.0721	.0708	.0694	.0681