Lecture 7

Continuous Random Variables

Today's Updates / Reminders

- Course notes packs are in stock at the bookstore!
- Homework 3 is due today.
- Lab 3 using R is due Friday. Use scientific notation for the penultimate question. Use ALL your decimals for the last question. Don't round intermediately.
- Homework 4 opens today.
- Stat Hub is open for business. Free help/tutoring. SAS1101, M-Th, 5:30-8:00pm

Mean and Variance of a Continuous R.V.

- Mean
  - For discrete: $\mu_x = E(x) = \sum[x \cdot P(x)]$
  - For continuous: $\mu_x = E(x) = \int x \cdot f(x)\, dx$
  - Note this mean is also called expected value.
- Variance
  - For discrete: $\sigma_x^2 = \sum(x - \mu_x)^2 \cdot P(x)$ OR $[\sum x^2 \cdot P(x)] - \mu^2$
  - For continuous: $\sigma_x^2 = \int_{-\infty}^{\infty}(x - \mu_x)^2 \cdot f(x)\, dx$ OR $[\int_{-\infty}^{\infty} x^2 \cdot f(x) dx] - \mu^2$
  - Note again that standard deviation, $\sigma_x = \sqrt{\sigma_x^2}$

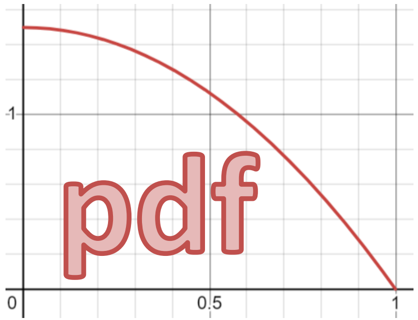
$$F'(x) = f(x)$$

$$F(x) = \int f(x)\,dx$$

Expected Value

$$f(x) = pdf$$

$$F(x) = cdf$$

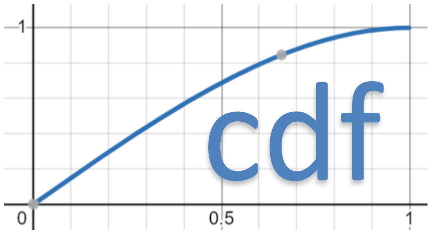


- The pdf of weekly gravel sales X is:

$$f(x) = \begin{cases} \frac{3}{2}(1 - x^2), & 0 \le x \le 1 \\ 0, & otherwise \end{cases}$$

- Therefore, we end up with E(x) as:

$$E(x) = \int_{-\infty}^{\infty} x \cdot f(x) dx = \int_0^1 x \cdot \frac{3}{2}(1 - x^2) dx$$



Variance

- For variance, we can use the “shortcut” method (see Lec5, slide 14):

$$V(X) = E(X^2) - [E(X)]^2$$

$$E(X^2) = \int_{-\infty}^{\infty} x^2 \cdot f(x) dx = \int_{0}^{1} x^2 \cdot \frac{3}{2}(1 - x^2) dx$$

But then don’t forget to subtract $[E(X)]^2$!!

One More Review Question

- Given the distribution function of a continuous variable, $f(x) = \frac{4}{27}x\left(x^2 - \frac{9}{2}x + 6\right), 0 \leq x \leq 3$, find P(X > 1.5)

- Given E(X) = 1.7, find the standard deviation.

The Normal Distribution

- This is the first day of the rest of your life, right here
- A continuous rv X is said to have a **normal distribution** with parameters μ and σ, where -∞ < μ < ∞ and 0 < σ if the pdf of X is:

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

Simpler Representation

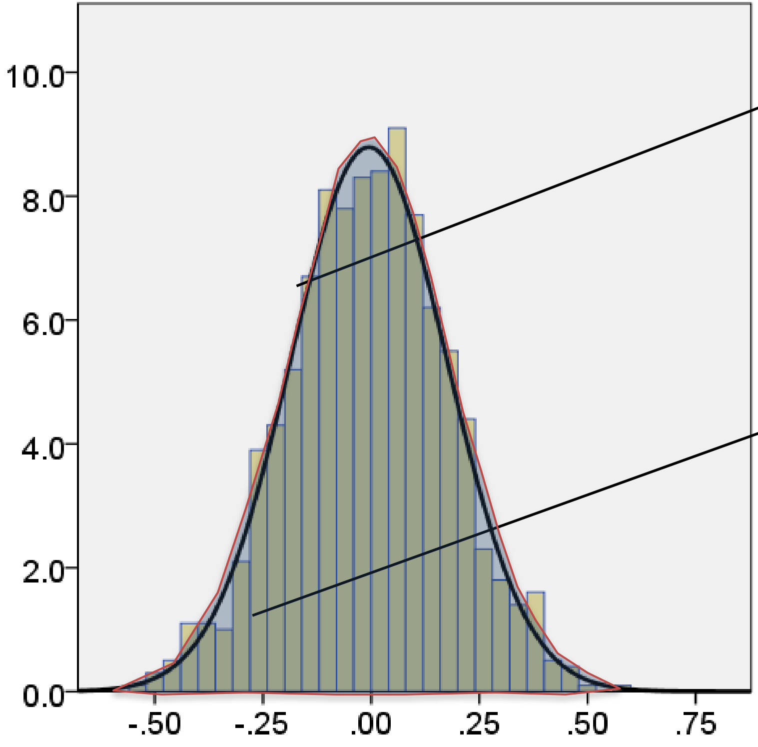
The mathematical expression for the height of the normal distribution is:

Gauss (1777-1855)

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

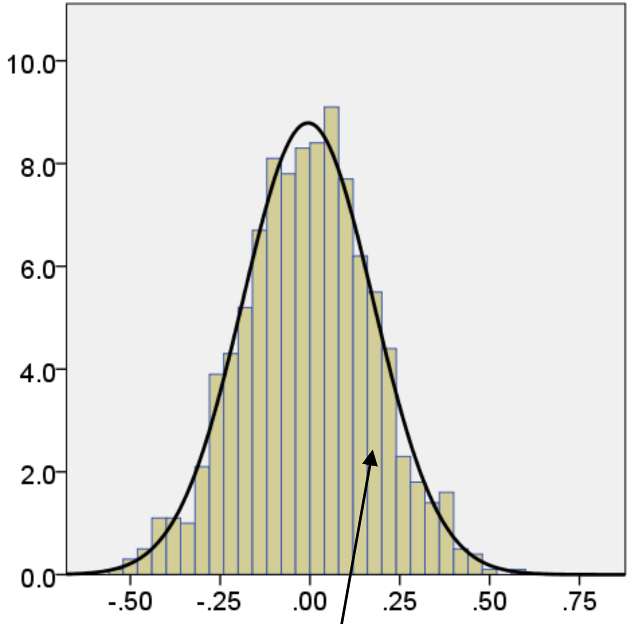
Given the values of the mean and the standard deviation, this is a function of x.

Density Curves and Normal Distribution



Density Curve: It is obtained by fitting a smooth curve to the histogram.
It gives the overall pattern of the data but ignores minor irregularities.
Area underneath equals 1 !!

How do we get the density curve?



Density curve → smooth approximation to histogram

How did we build the histogram?
We took a sample of size **n (n = #observations)**, and for each unit we measure a response variable. Then, we divided the x axis in bins and plotted the counts.

$$\sum_{b=1}^{\#bins} counts_b = n$$

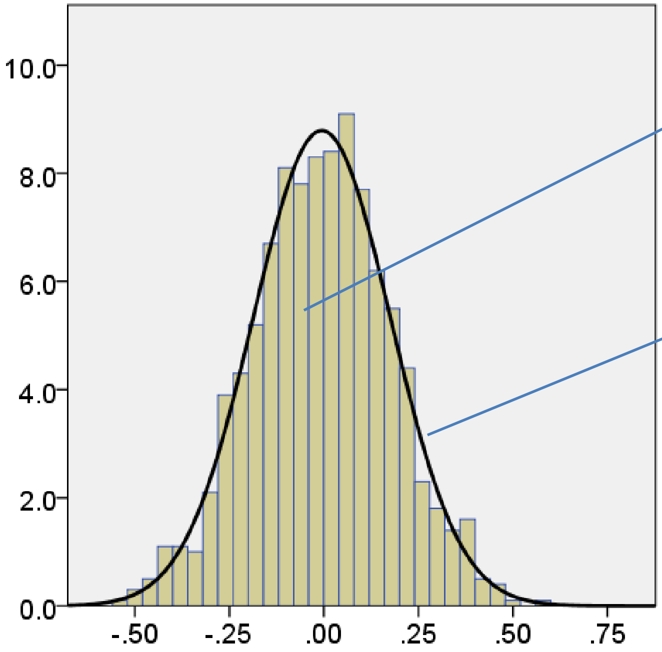
$$\sum_{b=1}^{\#bins} \frac{counts_b}{n} = 1$$

$$\sum_{b=1}^{\#bins} p = 1$$

p is related to the area of each of the bars

If we take a very large sample (**n** large), and a large number of bins, we get a smooth curve which area underneath equals 1 → **Density curve.**

Histogram to Density Curve



Suppose the variable we are interested is X (response)

Histogram (actual observations=sample distribution)
Mean → $\bar{x}$ → **Statistic**
Standard deviation → s

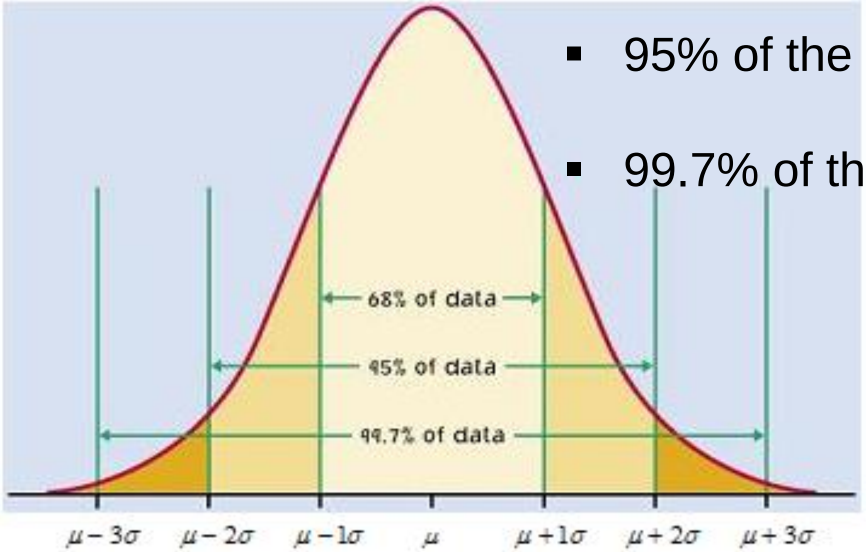
Density curve (idealized distribution=population distribution if the population is very large)
Mean → μ → **Parameter**
Standard deviation → σ

$\bar{x}$; s → mean and standard deviation of the sample dist

μ ; σ → mean and standard deviation of the population dist

The Empirical Rule

- 68% of the observations fall within σ of the mean μ
- 95% of the observations fall within 2σ of the mean μ
- 99.7% of the observations fall within 3σ of the mean μ



$P(\mu-1\sigma < X < \mu+1\sigma) = 0.68$

$P(\mu-2\sigma < X < \mu+2\sigma) = 0.95$

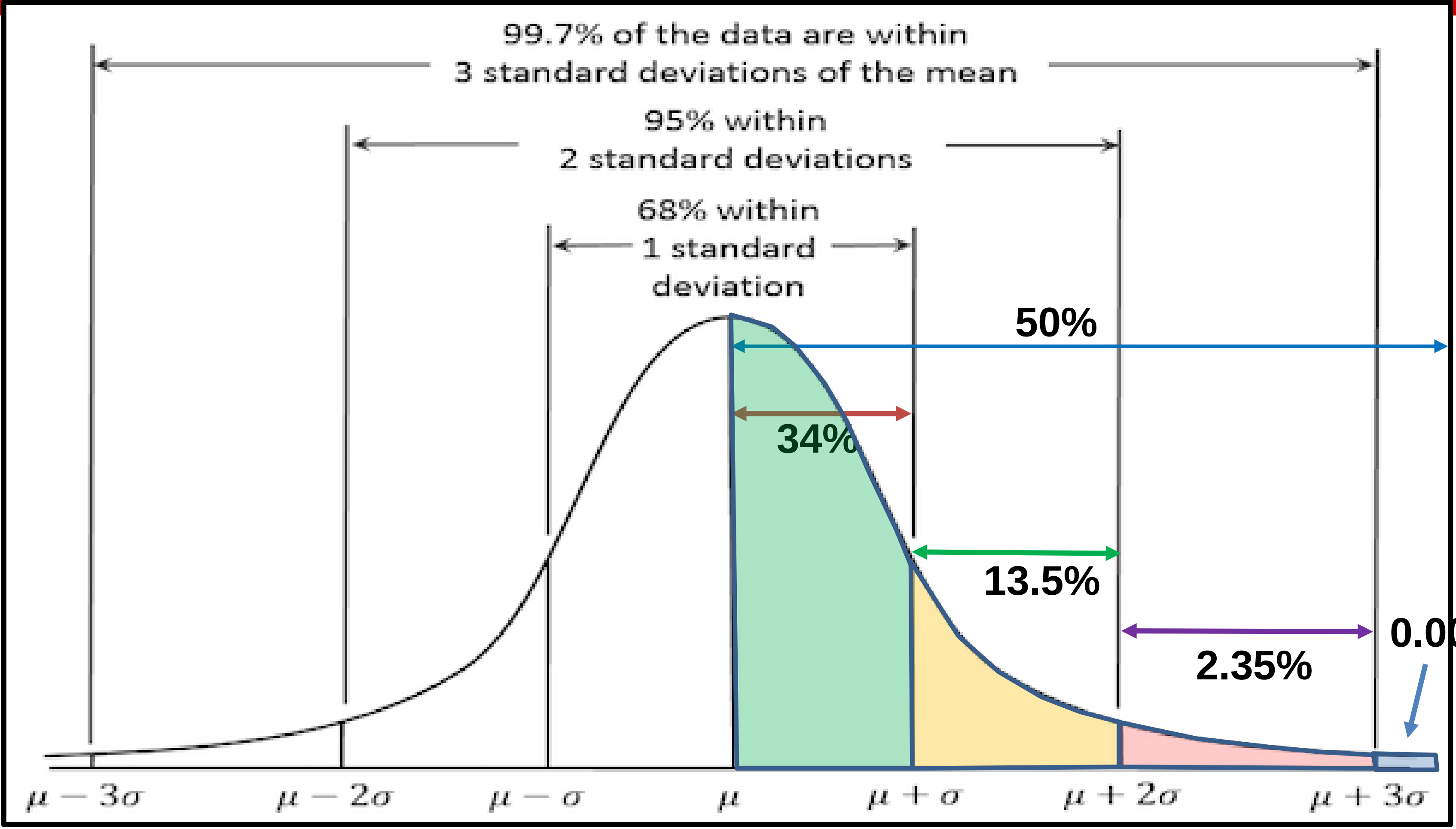
$P(\mu-3\sigma < X < \mu+3\sigma) = 0.997$

Remember: The whole area under the normal distribution =1

P= probability;

The area under the curve gives the probability of an observation falling between certain values of the variable.

Area = Probability!!

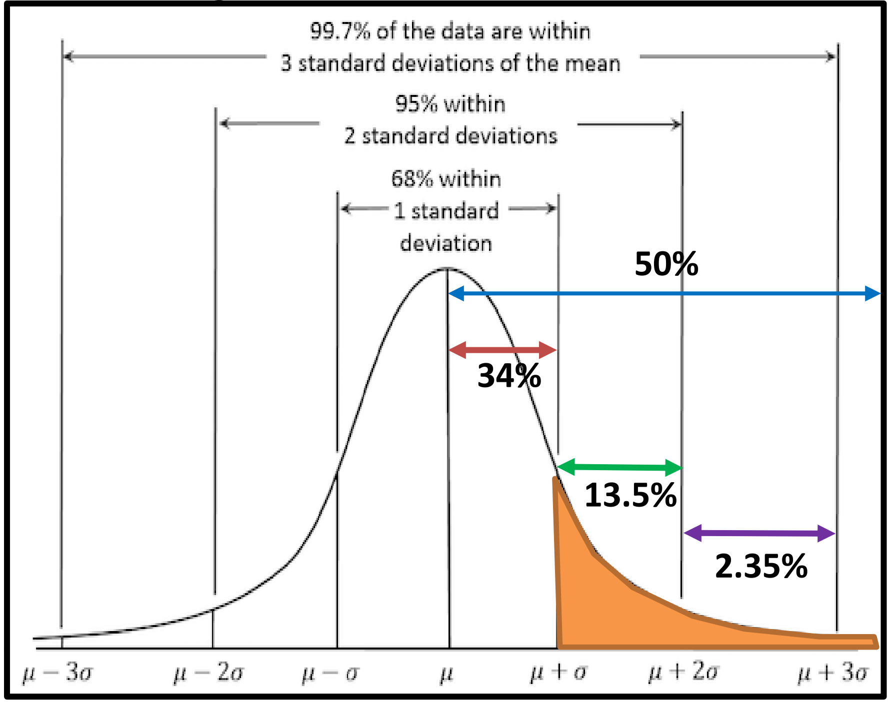

99.7% of the data are within
3 standard deviations of the mean
95% within
2 standard deviations
68% within
1 standard
deviation
50%
34%
13.5%
2.35%
0.0015
$\mu - 3\sigma$
$\mu - 2\sigma$
$\mu - \sigma$
μ
$\mu + \sigma$
$\mu + 2\sigma$
$\mu + 3\sigma$

Example

- Many states assess the skills of their students in various grades. One program is the NAEP which measures reading skills of 12th grade students. In a recent year the national mean score was 288 and the standard deviation was 38. Assuming that the scores are normally distributed with N(288,38),
- Use the 68-95-99.7 rule to give a range of scores that includes 95% of these students.
- If a total of 5674 12th grade students took the NAEP test, how many do you expect to have scores between 212 and 364?

Example (continued)

- If a total of 5674 12th grade students took the NAEP test, how many do you expect to score higher than 326? Remember: N(288,38)



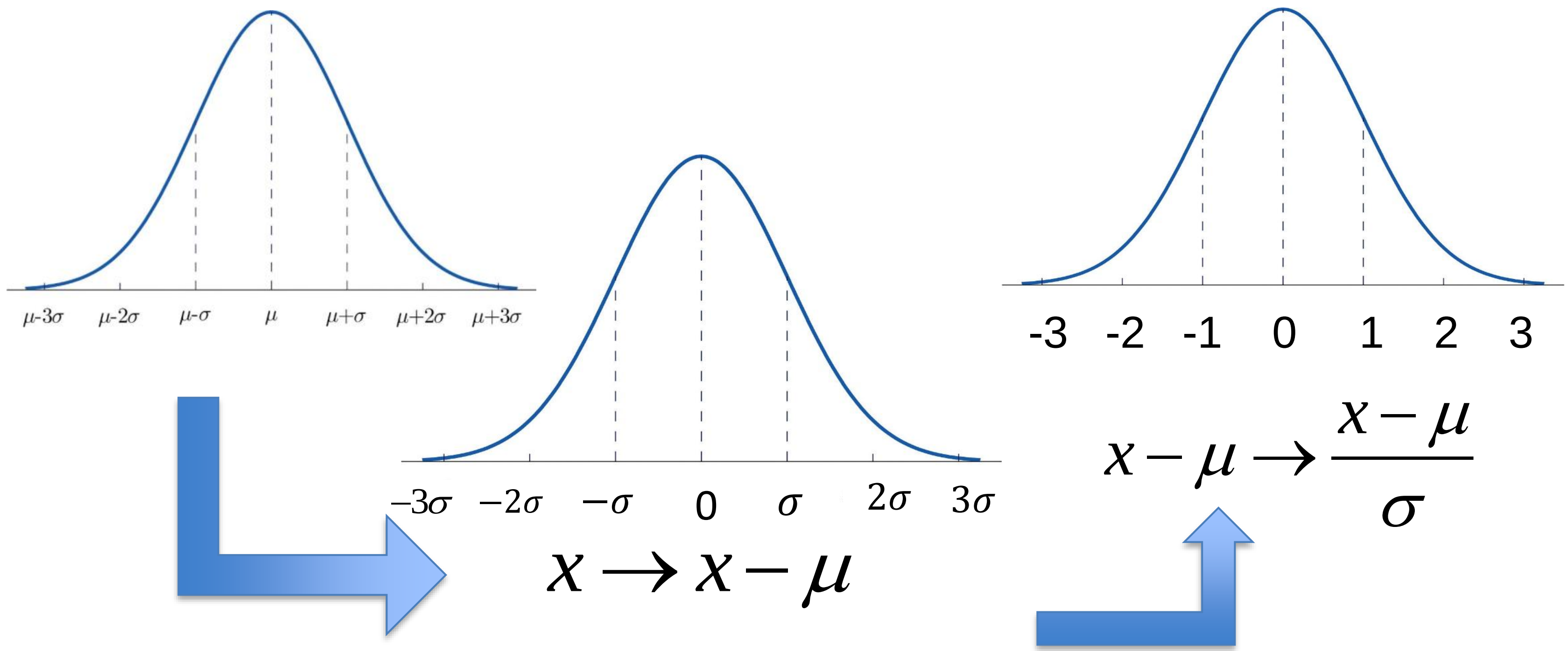
Standardizing the Variable

- Suppose that we want to compare two different normal distributions, how do we do that?

- Example: A person earns $25 an hour in state A. Another person earns $45 an hour in state B. You'd think the person in state B has better salary than the person in state A. However, if the distributions of salary per hour are normal in both states, but for state A we have N(22, 15) and for state B the distribution is N(50, 20). Which person has actually a better salary?

- To compare two distributions based on different measures we need to standardize the variable!!

Graphically

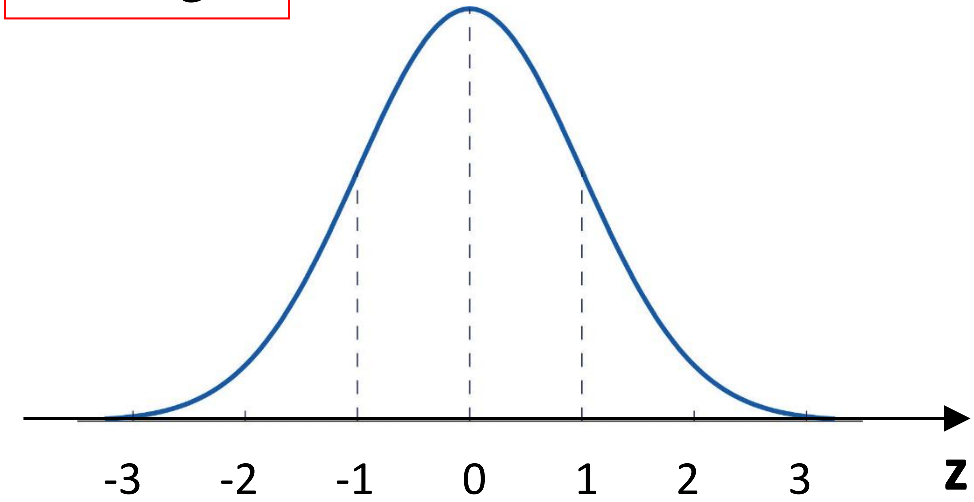


$$x \to x - \mu$$

$$x - \mu \to \frac{x - \mu}{\sigma}$$

Z

If we define $$z = \frac{x - \mu}{\sigma}$$ → z-score or standardized value



N(0,1)= Standard Normal Distribution

Why Standardize?

- A z-score tells us how many standard deviations the original observation falls from the mean.
  - $z>0$ means the observation is larger than the mean.
  - $z<0$ means the observation is smaller than the mean.
  - $z=1.3$ means the observation is more than one standard deviation above the mean.
- z-scores are useful to compare scores based on different measures.
- z-scores have a standard normal distribution N(0,1) associated to them, and there is a table (z-Table or Table A.3 at the end of textbook) which lets us find the probability for any given value of z. Also links to them on my website… negative Z, positive Z.

Example: Who has the higher score?

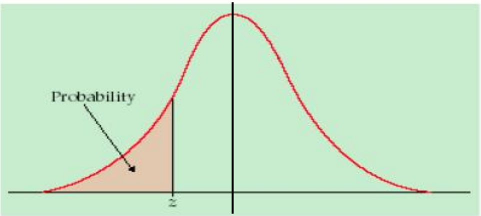
- Jacob scores 19 on the ACT. The ACT scores are roughly Normal with mean of 20.1 and standard deviation of 6.4.

- Andrew scores 995 on the SAT. The SAT scores are roughly Normal with a mean of 1040 and standard deviation of 224.

student	Jacob (ACT) N(20.1,6.4)	Andrew (SAT) N(1040,224)
Score (x)	19	995
Mean (μ)	20.1	1040
Standard deviation (σ)	6.4	224
Z-score (x-μ)/σ	(19-20.1)/6.4 = -0.172	(995-1040)/224 = -0.201

$P(Z < z_0) = ?$ → Z-Table (Table A.3 in book) of Standard Normal Probabilities

Table entry for z is the area under the standard normal curve to the left of z.



Important: This table only gives areas to the left of the z-score!!

Z-score→ margins
Probability→ inside

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
−1.3	.0968	.0951	.0934	.0918	.0901	.0885	.0869	.0853	.0838	.0823
−1.2	.1151	.1131	.1112	.1093	.1075	.1056	.1038	.1020	.1003	.0985
−1.1	.1357	.1335	.1314	.1292	.1271	.1251	.1230	.1210	.1190	.1170
−1.0	.1587	.1562	.1539	.1515	.1492	.1469	.1446	.1423	.1401	.1379
−0.9	.1841	.1814	.1788	.1762	.1736	.1711	.1685	.1660	.1635	.1611
−0.8	.2119	.2090	.2061	.2033	.2005	.1977	.1949	.1922	.1894	.1867
−0.7	.2420	.2389	.2358	.2327	.2296	.2266	.2236	.2206	.2177	.2148
−0.6	.2743	.2709	.2676	.2643	.2611	.2578	.2546	.2514	.2483	.2451
−0.5	.3085	.3050	.3015	.2981	.2946	.2912	.2877	.2843	.2810	.2776
−0.4	.3446	.3409	.3372	.3336	.3300	.3264	.3228	.3192	.3156	.3121
−0.3	.3821	.3783	.3745	.3707	.3669	.3632	.3594	.3557	.3520	.3483
−0.2	.4207	.4168	.4129	.4090	.4052	.4013	.3974	.3936	.3897	.3859
−0.1	.4602	.4562	.4522	.4483	.4443	.4404	.4364	.4325	.4286	.4247
−0.0	.5000	.4960	.4920	.4880	.4840	.4801	.4761	.4721	.4681	.4641

$P(Z<-2.56) = 0.0052$

$P(Z<-1.67) = 0.0475$

$P(Z<-0.42) = 0.3372$

Introduction to the practice of statistics. Moore, McCabe, Craig

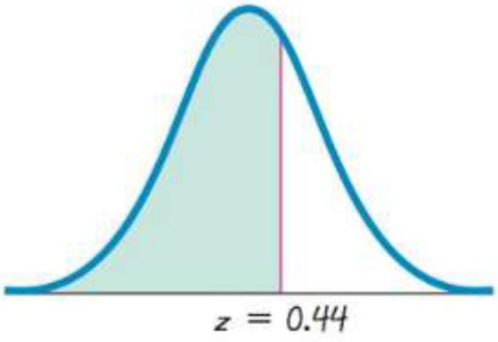
The Standard Normal Distribution

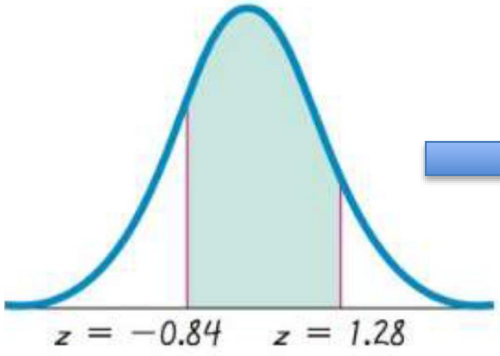
- Repeat of slide 5 with minor edits!
- A normal distribution with parameter values μ=0 and σ=1 is said to have a **standard normal distribution** with parameters μ and σ, where -∞ < μ < ∞ and 0 < σ if the pdf of X is:

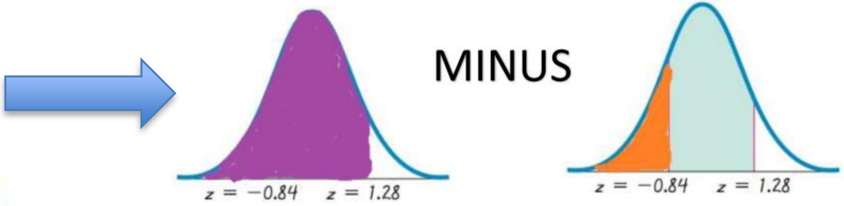
$$f(z; 0, 1) = \frac{1}{\sqrt{2\pi}} e^{\frac{-z^2}{2}}, -\infty < z < \infty$$

Normal Distribution

- There is no closed form integral of the normal probability function. We must use numerical methods (series). For the normal curve, we will use technology or software. Sheets, MATLAB, R, etc., ….







[Google Sheets]

[MATLAB]
normcdf, norminv

[R]
pnorm, qnorm

[Desmos]

[SALT]

Finding Probability ("Forward")

- Suppose the force acting on a column that helps to support a building is a normally distributed random variable X with mean value of 15.0 kips and standard deviation 1.25 kips. Compute the following probabilities.
  - $P(X \le 17.5)$
  - $P(X \ge 10)$
  - $P(14 < X \le 18)$

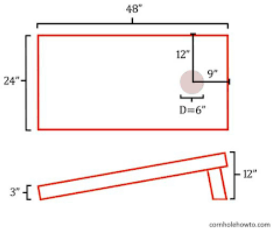
Percentiles (“Backward”)

- The heights of white pine trees are normally distributed with a mean of 95.2 feet and a standard deviation of 14.7 feet. What are the 40^{th} and 60^{th} percentiles?

Negative Z Tables
Positive Z Tables

Critical Value (z_α)

- You work for the American Cornhole Association (it's a thing!) and you're in the quality control department.
- The diameter of a hole on a cornhole board is 6" ± ¼".
- If we reject more than 5% of our parts (α), we will create operational inefficiencies (reduction in throughput yield and all sorts of other 6Sigma malfeasance)
- What value of z will cause us to have a 5% rejection rate? Keep in mind, parts can be oversize or undersize.
- What is an acceptable σ for my process?



Chapter 4.4 and 4.5

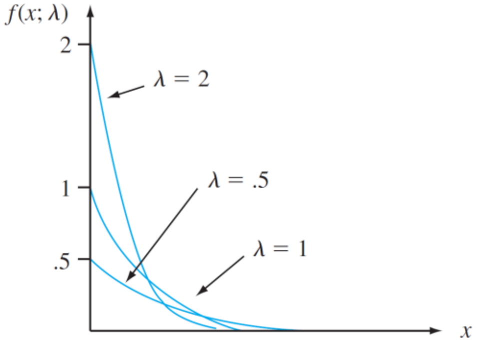
Other Distributions

The Exponential Distribution

$$f(x,\lambda) = \begin{cases} \lambda e^{-\lambda x}, & x \geq 0 \\ 0, & otherwise \end{cases}$$

$$F(x,\lambda) = \begin{cases} 0, & x < 0 \\ 1 - e^{-\lambda x}, & x \geq 0 \end{cases}$$

- This is closely related to the Poisson distribution.
- Mean, $\mu=1/\lambda$; standard deviation, $\sigma=1/\lambda$



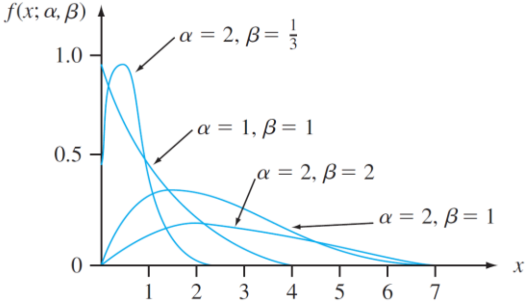
Exponential density curves

The Gamma Distribution

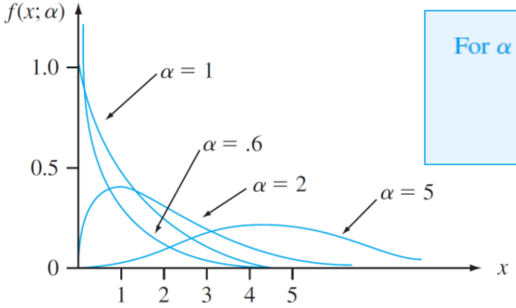
$$f(x;\alpha,\beta) = \begin{cases} \frac{1}{\beta^{\alpha}\Gamma(\alpha)} x^{\alpha-1} e^{-x/\beta}, & x \geq 0 \\ 0, & otherwise \end{cases}$$

Mean: $\alpha\beta$
Variance: $\alpha\beta^2$

- Used to model degradation, such as creep, corrosion, wear.
- Fun fact: if α=1 and β=1/λ, then it's the exponential distribution.



Gamma density curves



standard gamma density curves

For $\alpha > 0$, the **gamma function** $\Gamma(\alpha)$ is defined by

$$\Gamma(\alpha) = \int_0^{\infty} x^{\alpha-1}e^{-x}\,dx \qquad (4.6)$$

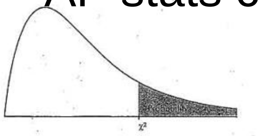
- The most important properties of the gamma function are the following:
- **1.** For any $\alpha > 1$, $\Gamma(\alpha)$= (α – 1) • $\Gamma(\alpha)$(α – 1) [via integration by parts]
- **2.** For any positive integer, *n*, $\Gamma(n)$= (n – 1)!
- **3.** $\Gamma(\frac{1}{2}) = \sqrt{\pi}$

The Chi-Squared Distribution

- Widely used distribution for comparing two categorical variables.

$$f(x;\upsilon) = \begin{cases} \frac{1}{2^{\frac{\upsilon}{2}}\Gamma(\frac{\upsilon}{2})} x^{(\upsilon/2)-1} e^{-x/2}, & x < 0 \\ 0, & x < 0 \end{cases}$$

- The parameter υ is called the number of degrees of freedom. The symbol X^2 is often used in place of chi-squared.
- There are various tests under chi-squared, many of which are included in an AP stats class. Test for independence, homogeneity, goodness of fit

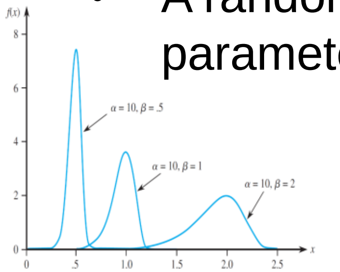


χ^2 CRITICAL VALUES

df	Tail probability p										
	.25	.20	.15	.10	.05	.025	.02	.01	.005	.0025	.001
1	1.32	1.64	2.07	2.71	3.84	5.02	5.41	6.63	7.88	9.14	10.83
2	2.77	3.22	3.79	4.61	5.99	7.38	7.82	9.21	10.60	11.98	13.82
3	4.11	4.64	5.32	6.25	7.81	9.35	9.84	11.34	12.84	14.32	16.27
4	5.39	5.99	6.74	7.78	9.49	11.14	11.67	13.28	14.86	16.42	18.47
5	6.63	7.29	8.12	9.24	11.07	12.83	13.39	15.09	16.75	18.39	20.51
6	7.84	8.56	9.45	10.64	12.59	14.45	15.03	16.81	18.55	20.25	22.46
7	9.04	9.80	10.75	12.02	14.07	16.01	16.62	18.48	20.28	22.04	24.32

	PASS SCREENING	DID NOT PASS SCREENING	TOTAL
WHITE	250	50	300
MINORITY	80	40	120
TOTAL	330	90	420

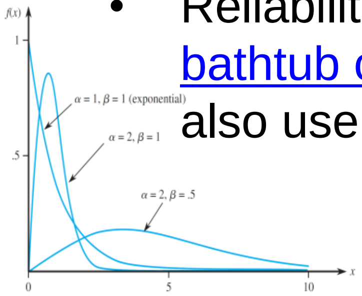
The Weibull Distribution

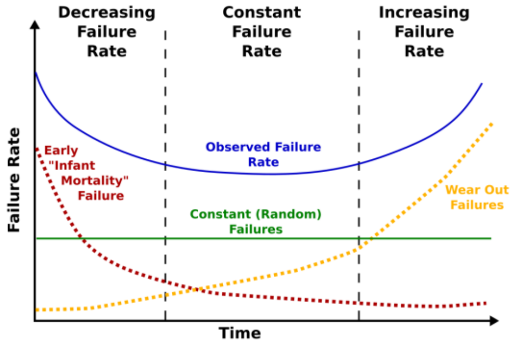


- A random variable X is said to have a Weibull distribution with shape parameter α and scale parameter β (α>0, β>0) if the pdf of X is:

$$f(x;\alpha,\beta) = \begin{cases} \frac{\alpha}{\beta^{\alpha}} x^{\alpha-1} e^{-(x/\beta)^{\alpha}}, & x \geq 0 \\ 0, & x < 0 \end{cases}$$

- When α=1, the pdf reduces to the exponential (with λ=1/β), so the exponential is special case of both gamma and Weibull. However, there are gamma that are not Weibull and vice versa.
- Reliability engineering uses the Weibull to create the [bathtub curve](). But in fact, financial profit models can also use the bathtub curve.





Mean: $\beta\Gamma\left(1+\frac{1}{\alpha}\right)$

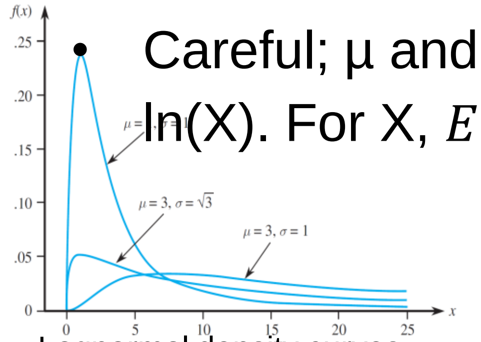
Variance: $\beta^2\left\{\Gamma\left(1+\frac{2}{\alpha}\right)-\left[\Gamma\left(1+\frac{1}{\alpha}\right)\right]^2\right\}$

The Lognormal Distribution

- A nonnegative rv X is said to have lognormal distribution if the rv Y = ln(X) has a normal distribution. The resulting pdf of a lognormal rv when ln(X) is normally distributed with parameters μ and σ is:

$$f(x;\mu,\sigma) = \begin{cases} \frac{1}{\sqrt{2\pi}\sigma x} e^{-[\ln(x)-\mu]^2/(2\sigma^2)}, x \geq 0 \\ 0, \qquad x < 0 \end{cases}$$

- Careful; μ and σ are not mean and standard deviation of X, but of ln(X). For X, $E(X) = e^{\mu+\sigma^2/2}$ and $V(X) = e^{2\mu+\sigma^2} \cdot (e^{\sigma^2} - 1)$



Lognormal density curves

The Beta Distribution

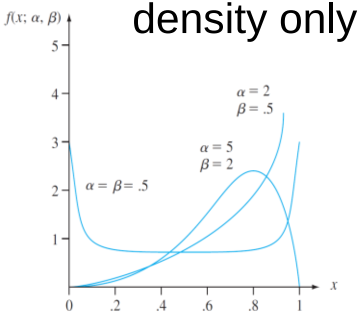
Mean: $A + (B - A) \cdot \frac{\alpha}{\alpha+\beta}$

Variance: $\frac{(B-A)^2\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$

- A random variable is said to have a beta distribution with parameters α, β (both positive), A, and B if the pdf of X is:

$$f(x;\alpha,\beta,A,B) = \begin{cases} \frac{1}{B-A} \cdot \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\cdot\Gamma(\beta)} \left(\frac{x-A}{B-A}\right)^{\alpha-1} \left(\frac{B-x}{B-A}\right)^{\beta-1}, & A \le x \le B \\ 0, & otherwise \end{cases}$$

- So far, every distribution has had positive density over an infinite interval (with the exception of the uniform distribution). They also all rapidly decrease to zero beyond a few standard deviations from the mean. The beta distribution provides positive density only for an X in an interval of finite length.



Standard beta density curves

- Unless α and β are integers, integration of the pdf to calculate probabilities is difficult.
- Commonly used to model variation in proportion of a quantity occurring in different samples.
  - Ex., proportion of a 24-hour day that you sleep