Lecture 8

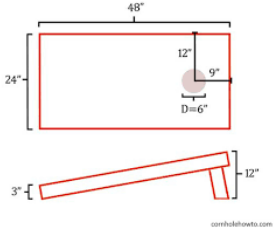
Joint Probability Distributions

Today's Updates / Reminders

- Lab 4 using Excel opens today. Very simple and due next Friday.
- Homework 4 is due the next time we meet.
- One more lecture (Lec 9) and then we will have a review for our first test the next class. Exam review and answer key have been posted to Moodle and my website. And then the first test is the class after that.
  - You may bring a formula sheet, front and back, anything you want to write, but has to be handwritten, on 8½" x 11" size paper
  - Bring your student ID and calculator, graphing or otherwise.
  - There will be no tables provided, but also no need for them.

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

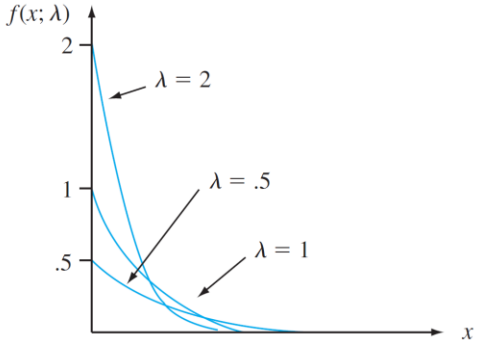
Other Distributions

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, μ=1/λ; standard deviation, σ=1/λ



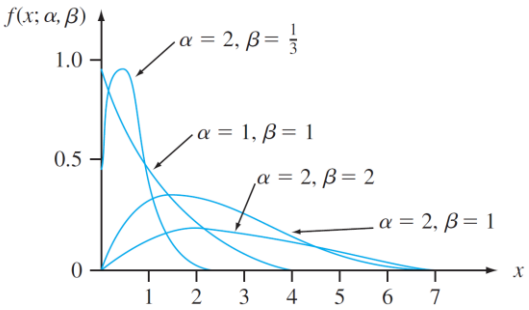
Exponential density curves

The Gamma Distribution

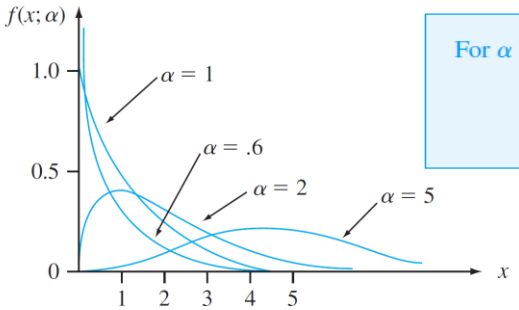
$$f(x;\alpha,\beta)=\begin{cases}\frac{1}{\beta^{\alpha}\Gamma(\alpha)}x^{\alpha-1}e^{-x/\beta}, & x\geq 0\\ 0, & otherwise\end{cases}$$

Mean: αβ
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

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

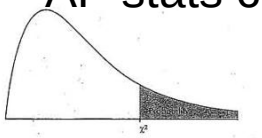
- The most important properties of the gamma function are the following:
- **1.** For any $\alpha > 1$, $\Gamma(\alpha)= (\alpha - 1)\cdot\Gamma(\alpha)(\alpha - 1)$ [via integration by parts]
- **2.** For any positive integer, *n*, $\Gamma(n)= (n - 1)!$
- **3.** $\Gamma(\frac{1}{2})=\sqrt{\pi}$

The Chi-Squared Distribution

- Widely used distribution for comparing two categorical variables.

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

- 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

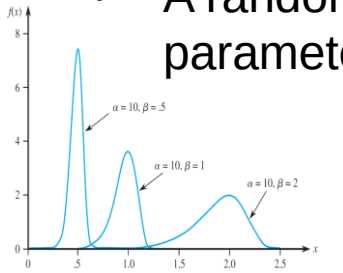


χ^2 CRITICAL VALUES

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

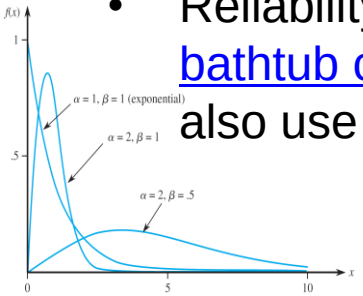
The Weibull Distribution



- 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:

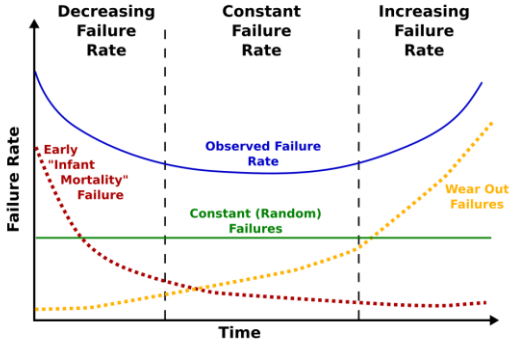
$$f(x;\alpha,\beta) = \begin{cases} \frac{\alpha}{\beta^{\alpha}} x^{\alpha-1} e^{-(x/\beta)^{\alpha}}, & x \geq 0 \\ 0, & x < 0 \end{cases}$$

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

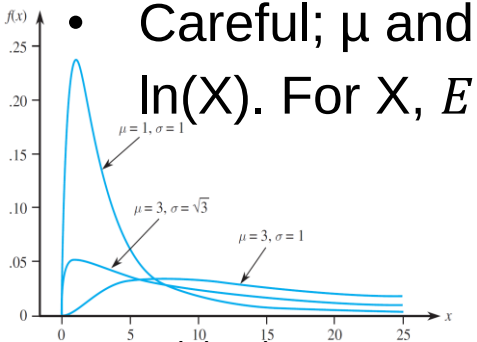


The Lognormal Distribution

- 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:

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

- 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

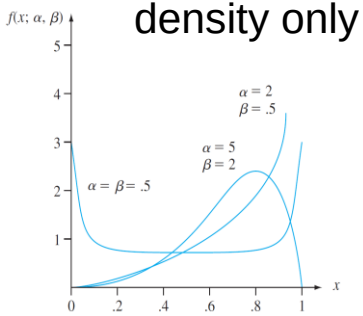
Mean: $A + (B - A) \cdot \frac{\alpha}{\alpha+\beta}$

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

Chapter 4.6

Probability Plots

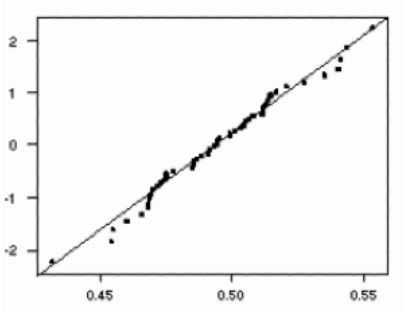
Normal Quantile (Q-Q) Plot

- Quartiles (4ths), Percentiles (100ths), Quantiles ("n"ths)
- Normal quantile plots provide a way to assess if a distribution is approximately normal. It is more sensitive than simply looking at a histogram.
  - Order the data from smallest to largest, x_1, x_2, …, x_n, where we use the notation x_i to indicate position
  - Compute the quantile → $q_i = \frac{i-0.5}{n}$
  - Find the z-score for each q_i and these are **expected** quantiles
  - Plot these z's against the data on the y-axis.

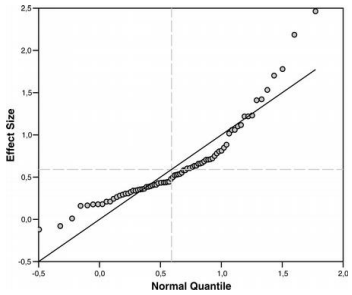
Google Sheets

More Details

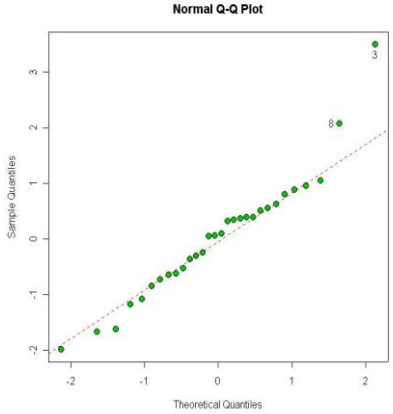
- If points lie close to a 45 degree straight line → the distribution is close to Normal. If few points are away from the overall pattern → those are outliers
- The slope of the line indicates the ratio of the standard deviation of your data to the standard deviation of a normal distribution.
  - Slope > 1, data is more spread out
  - Slope < 1, data is more tightly clustered.



Stat.yale.edu



Psycnet.apa.org



stat.wmich.edu

Appendix: Other Q-Q plotting positions

- If interested, there is some extra discussion on how to best find the plotting position.
- Other textbooks (and even some software like SAS) use an alternate method of calculating the quantiles. Their formula is:

$$q_i = \frac{i - 0.375}{n + 0.25}$$

Video showing this method

- And continuing in that article, they provide a table to test the correlation, r, of the best-fit line for a Q-Q plot. A screen snip of that table is provided on the next slide.

Looney, S. W., & Gulledge, T. R. (1985). Use of the Correlation Coefficient with Normal Probability Plots. *The American Statistician*, *39*(1), 75–79. https://doi.org/10.2307/2683917

Table 2. Empirical Percentage Points for Correlation Coefficient Test Based on Blom's Plotting Position

	Level													
n	.000	.005	.010	.025	.050	.100	.250	.500	.750	.900	.950	.975	.990	.995
3	.866	.867	.869	.872	.879	.891	.924	.966	.992	.999	.9997	.9999	1.000	1.000
4	.785	.813	.824	.846	.868	.894	.931	.958	.979	.992	.996	.998	.999	1.000
5	.729	.807	.826	.856	.880	.903	.934	.960	.977	.988	.992	.995	.997	.998
6	.686	.820	.838	.866	.888	.910	.939	.962	.977	.986	.990	.993	.996	.997
7	.651	.828	.850	.877	.898	.918	.944	.964	.978	.986	.990	.992	.995	.996
8	.623	.840	.861	.887	.906	.924	.948	.966	.978	.986	.990	.992	.994	.995
9	.599	.854	.871	.894	.912	.930	.952	.968	.980	.986	.990	.992	.994	.995
10	.578	.862	.879	.901	.918	.934	.954	.970	.980	.987	.990	.992	.994	.995
11	.560	.870	.886	.907	.923	.938	.957	.972	.981	.987	.990	.992	.994	.995
12	.544	.876	.892	.912	.928	.942	.960	.973	.982	.988	.990	.992	.994	.995
13	.529	.885	.899	.918	.932	.945	.962	.974	.983	.988	.991	.992	.994	.995
14	.516	.890	.905	.923	.935	.948	.964	.976	.984	.989	.991	.992	.994	.995
15	.504	.896	.910	.927	.939	.951	.965	.977	.984	.989	.991	.993	.994	.995
16	.493	.899	.913	.929	.941	.953	.967	.978	.985	.989	.991	.993	.994	.995
17	.483	.905	.917	.932	.944	.954	.968	.979	.986	.990	.992	.993	.994	.995
18	.473	.908	.920	.935	.946	.957	.970	.979	.986	.990	.992	.993	.9945	.9952
19	.465	.914	.924	.938	.949	.958	.971	.980	.987	.990	.992	.993	.9946	.9953
20	.457	.916	.926	.940	.951	.960	.972	.981	.987	.991	.992	.994	.9947	.9954
21	.449	.918	.930	.943	.952	.961	.973	.982	.987	.991	.993	.994	.995	.996
22	.442	.923	.933	.945	.954	.963	.974	.982	.988	.991	.993	.994	.995	.996
23	.435	.925	.935	.947	.956	.964	.975	.983	.988	.991	.993	.994	.995	.996
24	.429	.927	.937	.949	.957	.965	.976	.983	.988	.992	.993	.994	.995	.996
25	.422	.929	.939	.951	.959	.966	.976	.984	.989	.992	.993	.994	.995	.996
26	.417	.932	.941	.952	.960	.967	.977	.984	.989	.992	.993	.994	.995	.996
27	.411	.934	.943	.953	.961	.968	.978	.985	.989	.992	.994	.995	.9955	.9960
28	.406	.936	.944	.955	.962	.969	.978	.985	.990	.992	.994	.995	.9955	.9960
29	.401	.939	.946	.956	.963	.970	.979	.985	.990	.993	.994	.995	.9956	.9961
30	.396	.939	.947	.957	.964	.971	.979	.986	.990	.993	.994	.995	.9957	.9962
31	.392	.942	.950	.958	.965	.972	.980	.986	.990	.993	.994	.995	.9957	.9962
32	.387	.943	.950	.959	.966	.972	.980	.987	.991	.993	.994	.995	.9958	.9963
33	.383	.944	.951	.961	.967	.973	.981	.987	.991	.993	.994	.995	.9959	.9963
34	.379	.946	.953	.962	.968	.974	.981	.987	.991	.993	.994	.995	.996	.997
35	.375	.947	.954	.962	.969	.974	.982	.987	.991	.994	.9945	.9953	.996	.997
36	.371	.948	.955	.963	.969	.975	.982	.988	.991	.994	.9946	.9954	.996	.997
37	.368	.950	.956	.964	.970	.976	.983	.988	.991	.994	.995	.9955	.9962	.997

Looney, S. W., & Gulledge, T. R. (1985). Use of the Correlation Coefficient with Normal Probability Plots. *The American Statistician*, *39*(1), 75–79. https://doi.org/10.2307/2683917

Chapter 5.3

Statistics and Their Distributions

Sampling Distribution

- Data comes in all shapes and sizes. Last class we talked about multiple possible distributions of data.
- In addition to the variation that occurs from one data point to the next (i.e., measuring the volume in cans of Sun Drop off the assembly line), there is also variation that occurs if we draw a sample of data from a population. Imagine randomly drawing 3 numbers from 0 to 10 from a hat and finding the mean. The numbers (data) in this situation are not changing. It is the sample that is changing.
- The latter is what is happening when we take a subset (sample) of items from a population. We will have variation because each subject varies. But we will also get variation of any statistic we calculate on our sample.

Sampling vs. Population Distribution

- We need to distinguish between two distributions:
  - Sampling Distribution: Is the probability distribution of the statistic. We calculate the statistic from the data (response variable) we collected from each sample.
  - Population Distribution: of a variable X is the distribution of its values for all members of the population. It is also the probability distribution of the variable when we choose one individual at random from the population. (Note: the parameter is the measure of center of the population distribution, usually the mean)
- Typical statistics that describe quantitative data or quantitative variables: sample mean, median, and standard deviation → We'll study the sample mean→ Statistic = Sample mean

Sampling

- Sometimes we can calculate an estimate for a population mean theoretically via probability.
  - You work at the Howling Cow Creamery, making Howling Cow Patties ($7), Howling Calf Patties ($5) and milkshakes ($6). The probability of each of these being made and sold follows:

Calf Patty	Milkshake	Cow Patty
$5	$6	$7
0.3	0.2	0.5

Find the expected value.

Deriving a Sampling Distribution

- Let's look at two different customers for these 3 products. Let X_1 and X_2 denote the customers.
- What are the possible arrangements of items purchased by these customers, assuming they pick one of the three?

What are the probabilities of these?

Calf Patty	Milkshake	Cow Patty
$5	$6	$7
0.3	0.2	0.5

	$p(x_1, x_2)$	$\bar{x}$	s^2
Cow-Cow			
Cow-Calf			2
Cow-Shake			0.5
Calf-Cow			
Calf-Calf			
Calf-Shake			
Shake-Cow			
Shake-Calf			
Shake-Shake			

Sampling Distribution of $\bar{X}$ and S^2

$\bar{x}$	\$5	\$5.50	\$6	\$6.50	\$7
$P_{\bar{X}}(\bar{x})$					

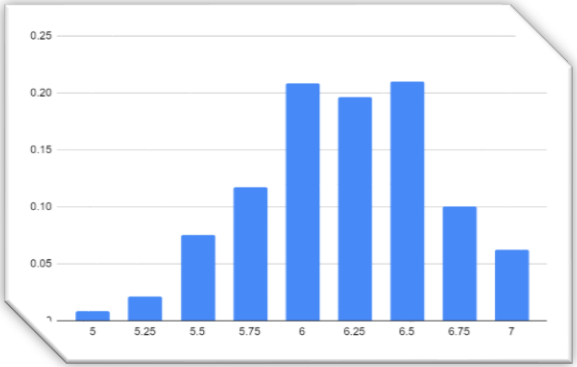
s^2	0	0.5	2
$P_{S^2}(s^2)$			

- Let's sketch a histogram of the mean using the original distribution and the distribution above.
- What's the expected value?

What if we looked at more customers?

- If we bumped it up to 4 customers, things get much more tedious… 3^4 arrangements of sales…

$\bar{x}$	\$5	\$5.25	\$5.50	\$5.75	\$6.00	\$6.25	\$6.50	\$6.75	\$7.00
$P_{\bar{X}}(\bar{x})$	0.0081	0.0216	0.0756	0.1176	0.2086	0.196	0.21	0.1	0.0625



- Note that expected value is still 6.2, but we get more resolution in the data.

Google Sheets

Sampling

- But is the probability of each of those items being sold a fixed number? Not likely. In some cases, we can know the probability of a situation (flipping a coin, rolling a die, getting a green stoplight).
- Sampling is a way to ascertain the mean. As you can easily imagine, we could just take the sales data for those 3 products over some time period.
  - And of course, realize that day to day has its own variability
  - And when we sample, we will also have variability.

Rolling Dice

- In fact, I surreptitiously did this when we did our dice rolling experiment.
- Note how I have a mean of each of your dice rolls. There were hundreds of rolls of the dice and I selected your 5 and averaged them.
- When the means are plotted, we can see the distribution of the sample means.

Jump Experiment

- This is actual data that I collected from students. We placed a piece of tape on the floor. They lined up their toes with the tape, stood flat-footed and jumped as far as they could. We measured the distance of their jump.
- Most students tried their best, but some were “too cool for school.” But also consider variability in gender, leg length, athletic conditioning, and even type of shoe/clothes you are wearing.
- How would you describe this data?
- In fact, let’s go create a normal Q-Q plot to see what that looks like.

Sampling Distributions and n

- Let's go see this in StatCrunch. Go to Moodle. Under Practice Resources. Click Applets, Sampling Distribution
- Try some of the different distribution types… uniform, right-skewed, etc.
- Click Compute!
- On the next screen, leave sample size as 2 and click to do it 5 times. Notice how a gray sample drops down from the distribution, 2 back to back.
- Now change sample size to 5 and try it again.
- Hopefully you are seeing that, as sample size increases, the means of the samples are closer to 25 every time and the spread between the means is decreasing. Note this MAY not happen, due to randomness.