

POL 602
Prof. Matthew Lebo
Week 7

Functions of Random Variables and Misc. Topics

[Note: we are returning to sections 3.9, 3.11 and 4.9 for the first part of this discussion]

The mean and standard deviation provide useful information about the distribution of a random variable, but they do not provide a unique characterization of the random variable.

In other words, many (an infinite number) distributions have the same means and standard deviations. Under some general conditions it is possible to use a numerical means of uniquely describing a distribution. First, however, we need to introduce some terms.

The k^{th} moment about the origin of a random variable Y is written as μ'_k and is defined: $\mu'_k = E(Y^k)$. Thus, the first moment about the origin is:

$$\mu'_1 = E(Y^1) = E(Y) = \mu$$

That is, the mean of the random variable. A second type of moment which we frequently employ is the *moment about the mean* of a random variable:

$$\mu_k = E\left[(Y - \mu)^k\right]$$

Thus we can think of variance as the second moment about the mean of a random variable:

$$\mu_2 = E\left[(Y - \mu)^2\right]$$

I should note that the 0th moment about the mean is 1, and the 1st moment about the mean is equal to zero. Also interesting are the third and fourth moments about the mean:

$$\mu_3 = E\left[(Y - \mu)^3\right] \equiv \textit{skewness}$$

$$\mu_4 = E\left[(Y - \mu)^4\right] \equiv \textit{kurtosis}$$

Where skewness is the familiar property of asymmetry in a distribution.

Kurtosis is a measure of whether a variable is peaked or flat relative to a normal distribution.

A random variable with a high kurtosis tends to have a distinct peak near the mean, declines rather rapidly, and has heavy tails.

A variable with low kurtosis tends to have a flat top near the mean rather than a sharp peak. A uniform distribution would be the extreme case.

A *moment-generating function* gives us a way to, as the book says, pack “all of the moments for a random variable into one expression.” A moment-generating function

$$m(t) = E(e^{ty})$$

exists for a random variable Y if there exists a positive constant b such that $m(t)$ is finite for $|t| \leq b$. Why is this true? We need the Taylor series expansion of e^x :

$$e^x = \sum_{i=0}^{\infty} \frac{x^i}{i!}$$

Using this result, we can rewrite e^{ty} :

$$e^{ty} = 1 + ty + \frac{(ty)^2}{2!} + \frac{(ty)^3}{3!} + \dots$$

The expected value of this series is thus (for discrete and continuous random variables):

$$E(e^{ty}) = 1 + t\mu'_1 + \frac{t^2}{2!}\mu'_2 + \frac{t^3}{3!}\mu'_3 + \dots$$

which is a function of all of the moments about the origin (proof for discrete case is in section 3.9 and continuous case in section 4.9).

There are two useful applications of the moment-generating function. First, if we know the function (and the moment generating functions for many common distributions are given inside the back cover of the textbook, for example), we can find any moment for Y:

$$\left. \frac{d^k}{dt^k} m(t) \right|_{t=0} = \mu'_k$$

In other words, we take the k^{th} derivative of $m(t)$ with respect to t and set $t=0$ to find the k th moment about the origin of Y.

The other application of the moment-generating function is to prove that a random variable has a particular probability distribution.

If the moment-generating functions of two random variables Y and Z are equal, then the two random variables must have the same probability distribution. We will now use this to show some useful results concerning the distribution of sums and products of random variables.

First, if Y_1, Y_2, \dots, Y_n are independent random variables with moment-generating functions $m_{Y_1}(t), m_{Y_2}(t), \dots, m_{Y_n}(t)$.

If $U = Y_1 + Y_2 + \dots + Y_n$, then the moment-generating function of U is:

$$m_U(t) = m_{Y_1}(t) \times m_{Y_2}(t) \times \dots \times m_{Y_n}(t)$$

In words, the moment-generating function of the sum of n random variables is equal to the product of the moment-generating functions of n random variables (this is easy to prove given the law of exponents that states: $e^{a+b} = e^a e^b$; see theorem 6.2 in the text).

This result can then be applied to the sums of several types of random variables to determine how their sum is distributed. Two results of this type are of particular note:

If Y_1, Y_2, \dots, Y_n are independently distributed Normal random variables with $E(Y_i) = \mu_i$ and $V(Y_i) = \sigma_i^2$ for $i=1, 2, \dots, n$ and if

$$U = \sum_{i=1}^n a_i Y_i$$

(that is, the sum of i normal random variables each multiplied by an arbitrary constant), then U is also a normally-distributed random variable with:

$$E(U) = \sum_{i=1}^n a_i \mu_i$$

$$V(U) = \sum_{i=1}^n a_i^2 \sigma_i^2$$

Finally, if Z_i is a random variable with a standard normal distribution (mean 0, std. dev. 1), then:

$$\sum_{i=1}^n Z_i^2 \sim \chi^2 \text{ with } \nu = n \text{ degrees of freedom}$$

In other words, the sum of n squared standard normal random variables is distributed chi-square with n degrees of freedom (parameter n). This result is useful in calculating a variety of goodness-of-fit statistics.

This is cool, no? If $n=1$, then a squared standard normal is a χ^2 distribution with 1 degree of freedom. Look at the two distributions and the areas under the curve to convince yourself.

The last “house-keeping” topic I want to cover before moving on the second half of the course is a theorem that may be familiar to prior students of statistics: Tchebysheff’s (Chebechev’s) Theorem.

In the very first class, we discussed the so-called empirical rule which stated that, for roughly Normal data, we could approximate the percentage of values or outcomes that would be found in certain ranges:

$$P(\mu \pm \sigma) \approx .68$$

$$P(\mu \pm 2\sigma) \approx .95$$

$$P(\mu \pm 3\sigma) \approx .997$$

What if our data are not approximately normal? Many of the distributions we have discussed look markedly different, such as the Gamma-type family of distributions for many values of alpha and beta.

Tchebysheff’s theorem provides a conservative pair of upper and lower bounds for any number of standard deviations about the mean:

$$P(|Y - \mu| < k\sigma) \geq 1 - \frac{1}{k^2}$$

This is proven in section 4.10.

How does this compare to the empirical rule for bell-shaped distributions?

$$P(\mu \pm \sigma) \geq 0$$

$$P(\mu \pm 2\sigma) \geq .75$$

$$P(\mu \pm 3\sigma) \geq .889$$

The usefulness of this theorem is apparent in cases when we have no knowledge about the particular form of the distribution of a random variable.

For example: assume that we have no idea how the number of executive orders issued by a U.S. President is distributed, but we have observed all Presidents to date and have determined that they issue a mean number of 20 orders in each term, with a standard deviation of 2 orders (this can be computed easily from a list of data).

What can we say about the probability that the next President will issue more than 16 but fewer than 24 executive orders?

First, how many standard deviations around the mean of 20 is the range from 16-24?

Since one s.d. is equal to 2, this range is 2 s.d.’s about the mean:

$$[20-2(2), 20+2(2)]=[16,24]. \text{ So } k = 2.$$

Applying Tchebysheff’s theorem:

$$1 - \left(\frac{1}{2}\right)^2 P(16 < Y < 24) = P(|Y - \mu| < 2\sigma) \geq 1 - \frac{1}{2^2} \geq .75$$

So, there is *at least* a .75 probability that the next President will issue between 16 and 24 executive orders.

Another example:

Suppose that W is a random variable with $\mu = 150$ and $\sigma = 20$. Is the following equation possible?

$$P(110 < W < 190) = 0.65$$

No: $P(110 < W < 190) = P(\mu - 2\sigma < W < \mu + 2\sigma)$.

According to Tchebysheff's theorem, the probability *must* be greater than $1 - \left(\frac{1}{2}\right)^2 = 0.75$.

So, the probability cannot be as low as 0.65.

Some Tchebysheff problems: 4.114, 4.115, 4.116, 4.119, 4.120.

Sampling Distributions and the Central Limit Theorem

We now have enough probability theory to begin the study of statistics. As noted in the beginning of this course, the dominant method (in terms of the assignment of probabilities and the logic of inference) that we will be considering is the relative frequency or *frequentist* logic of statistics.

Where applicable, however, we'll make brief mention of the differences and relative strengths and weaknesses of the Bayesian analog.

Above we discussed several results using functions of several random variables, $Y_1, Y_2 \dots Y_n$, which were assumed to be independent, but with a common distribution (although not necessarily an *identical* one).

When we take a sample of data from a population, we observe values of several random variables in this manner. Certain functions of these random variables can be used to make estimates or decisions about the unknown population parameters.

A *statistic*, in fact, can be defined as a function of the observable random variables in a sample and known constants. A statistic will thus always be a random variable itself.

For example, say we wish to estimate the population mean of a particular variable, such as age. What would we usually use to estimate such a thing? The sample mean seems to be a reasonable choice:

$$\bar{Y} = \frac{1}{n} \sum_{i=1}^n Y_i$$

In this case \bar{Y} , the sample mean, is a function only of the random variables $Y_1, Y_2 \dots Y_n$ and the constant n (the sample size). The random variable \bar{Y} is thus a *statistic*.

Since statistics are random variables themselves, it is possible (using tools we discussed briefly in the section on moment-generating functions) to derive their probability distributions.

We call these *sampling distributions*. The sampling distribution of a statistic gives us a theoretical model of the relative frequency histogram of the possible values of the statistic that we would observe in repeated sampling.

This will lead directly to ways to measure *goodness* of our estimates.

It is worth emphasizing this point. Ideally, we would take many repeated samples from our population in order to get a good idea of the distribution of the estimate of the parameter we are interested in.

Since this is usually not feasible (except in simulation studies; see below), what we do instead is theoretically determine how the sample statistic is distributed based on assumptions about how the random variables are distributed.

This allows us to model the goodness of our estimate without actually using repeated samples.

Perhaps the most common assumption about the random variables observed in a sample is that they are distributed independently normal.

This is actually quite reasonable given the prevalence of the normal distribution in the natural and social sciences.

In the example above (the sample mean as an estimator for the population mean), what if $Y_1, Y_2 \dots Y_n$ are a random sample of size n from a normal distribution with mean μ and variance σ^2 (that is, they are *identically independently distributed* or *i.i.d.*)?

It can then be shown (proof is Theorem 7.1 in the text) that

$$\bar{Y} = \frac{1}{n} \sum_{i=1}^n Y_i$$

is distributed normal with mean $\mu_{\bar{Y}} = \mu$ and variance $\sigma_{\bar{Y}}^2 = \frac{\sigma^2}{n}$.

In other words, the sum of n normal random variables with identical mean μ and variance σ^2 is a random variable with mean μ and variance $\frac{\sigma^2}{n}$.

This is the sampling distribution of \bar{Y} .

Question: what is there in these formulae that is under the researcher's control that could lead to a better estimate of the mean of several random variables i.i.d.?

Answer: the sample size, n .

Similar logic leads to a different result in the case of standard normal random variables. If $Y_1, Y_2 \dots Y_n$ are a random sample of size n from a normal distribution with mean μ and variance σ^2 (as before), we can then define

$$Z_i = \frac{Y_i - \mu}{\sigma}$$

as standard normal variables. Using the theorem (6.4) from the previous discussion on moment-generating functions, it can then be shown that

$$\sum_{i=1}^n Z_i^2 = \sum_{i=1}^n \left(\frac{Y_i - \mu}{\sigma} \right)^2$$

is distributed χ^2 with n degrees of freedom.

Chapter 7 in the text provides the proofs of these two results and introduces several others (which we will use later). When necessary, I will introduce a particular result. For

now, it is important mainly to understand that the sampling distributions for various functions of random variables can be determined theoretically through methods such as these.

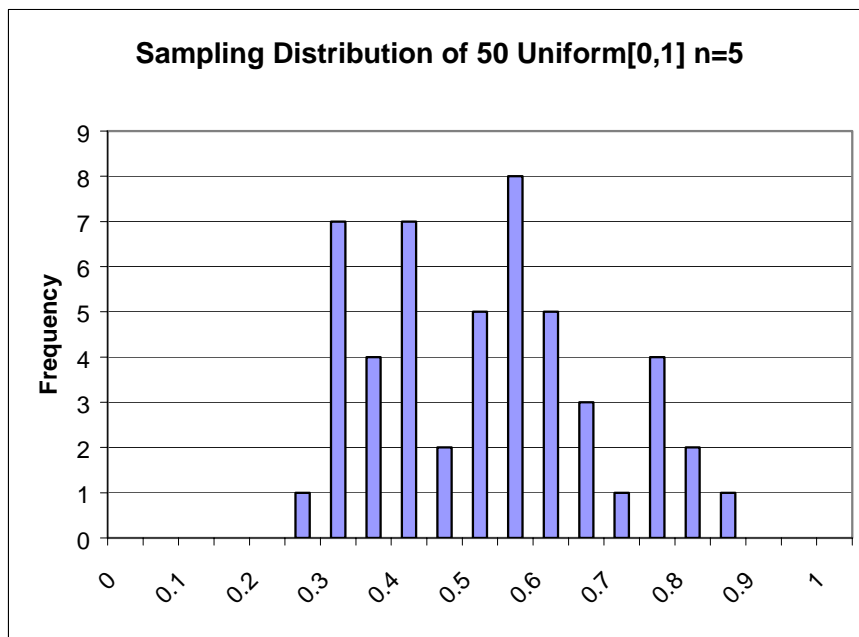
The Central Limit Theorem

We discussed in the preceding section that the sum of n normal random variables with identical mean μ and variance σ^2 is a random variable with mean μ and variance $\frac{\sigma^2}{n}$.

What if the random variables are i.i.d. but not normally-distributed?

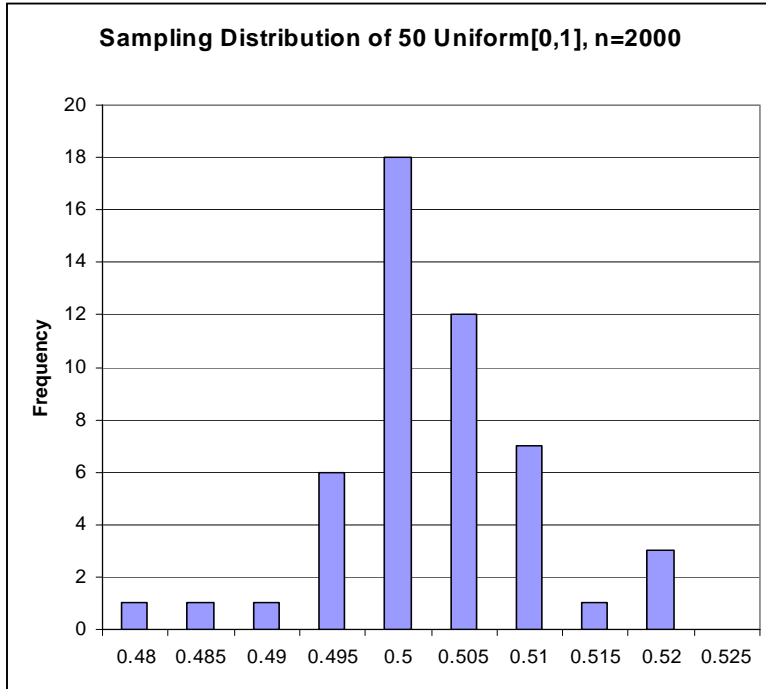
One strategy to find the sampling distribution of a function of them would be to use the method of moment-generating functions. Now, however, I want to introduce an important result that allows us to approximate the sampling distribution of interest.

We can use simulation techniques. Say $Y_1, Y_2, \dots, Y_n \sim \text{Uniform}[0,1]$ i.i.d. How is \bar{Y} , the sample mean, distributed? What should we expect the mean to be (.5)? What about the sampling distribution? If we have a sample size of 5, and we take 50 repeated samples, one possible empirical distribution is:



which has a mean of about .494 and a std. dev. of about .138.

If we increase the size of each of our 50 samples to 2000 (same distributions of uniform i.i.d. random variables) we get a sampling distribution of:



which has a mean of .5010, and std. dev. of .007.

It is obvious that the mean of the sampling distribution of \bar{Y} is equal to the population mean of 0.5. However, it is equally obvious that the sampling distribution is not distributed uniform. How does it look to be distributed? Approximately normal as the sample size gets bigger. This is the intuition behind the central limit theorem.

The CLT is important enough to go over it again. Another way, perhaps the simplest way, to state the central limit theorem is that:

The sum of a large number of i.i.d random variables will be approximately normal, regardless of their underlying distribution.

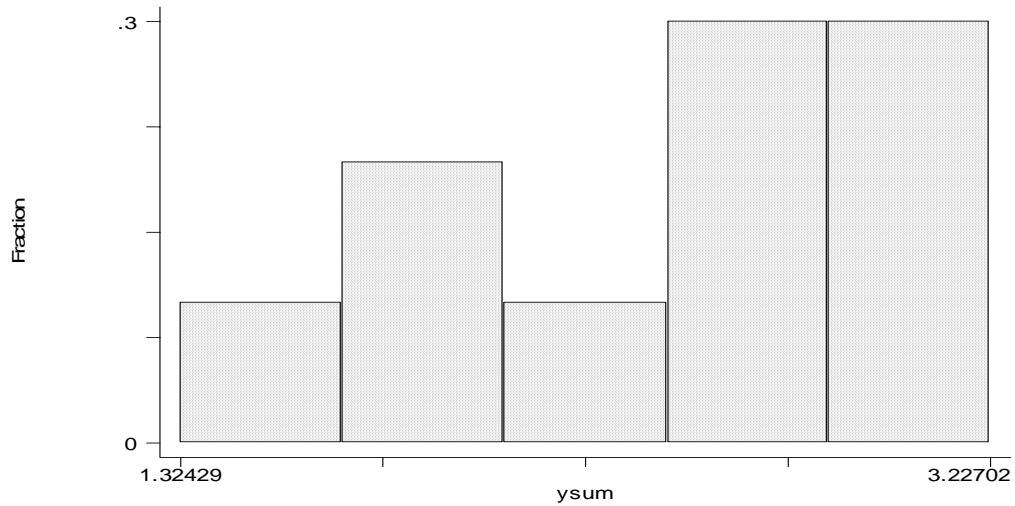
The larger the sample size and/or the larger the number of random variables added together, the closer the approximation will be to the Normal distribution. For example, the following slide shows the behavior of a random variable, Y_{sum} , which is the sum of five i.i.d. continuous random variables \sim Uniform[0,1].

In the first picture, the sample drawn for each random variable is only 10 ($n = 10$). The picture gives the empirical distribution for Y_{sum} using simulated pseudorandom variables (computer generated).

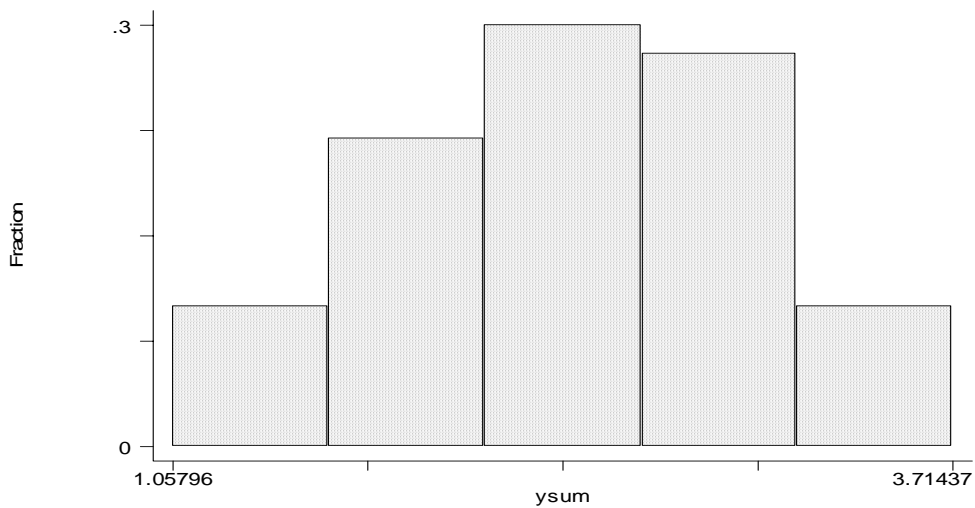
In the second picture, we are still adding together 5 uniform random variables. This time, however, the sample size for each random variable is 100. Note the beginning of a bell-shaped curve.

In the third picture, the same 5 random variables now have a sample size of 1000, and the histogram looks approximately normal.

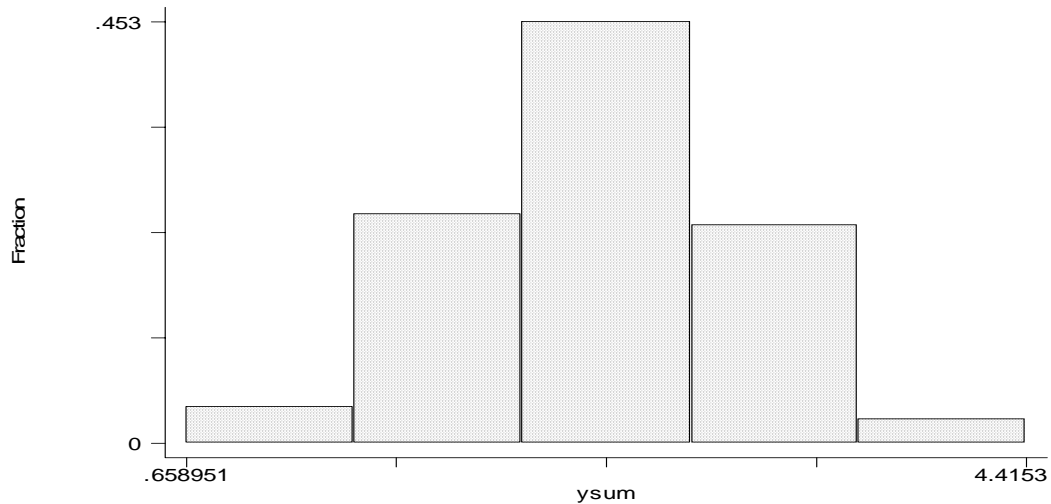
Question: What is the mean (expected value) of Ysum in theory for each case? What is it appear to be empirically? $Y_{sum} = \text{Sum of 5 random variables } \sim U[0,1]$
 $n = 10$



n = 100



$n = 1000$



Now, what if we hold the sample size constant, but increase the number of uniform random variables being added together?

To present these simulation results, let me quickly introduce a diagnostic tool for testing Normality: the normal Q-Q plot, which is a plot of the quantiles of an observed random variable against the normal distribution quantiles (all points along the diagonal indicates normality) [Q-Q = quantile-quantile].

In each picture, n is fixed at 10,000. In the first, the random variable is the sum of 2 Uniform[0,1]

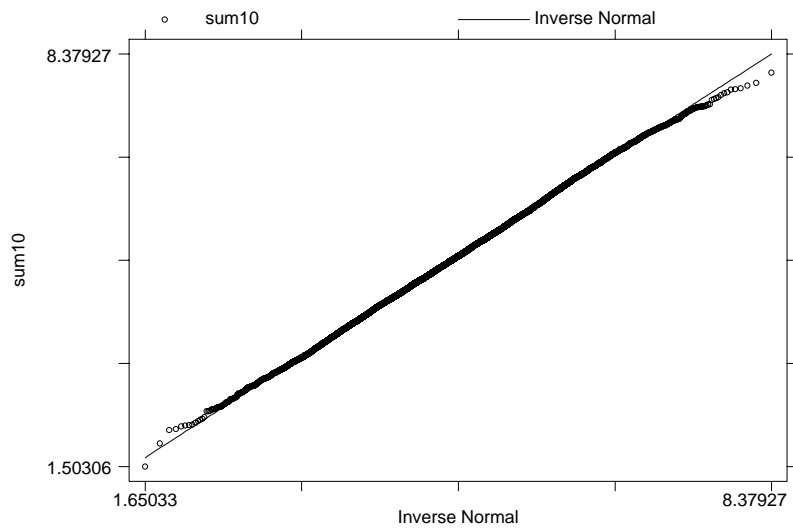
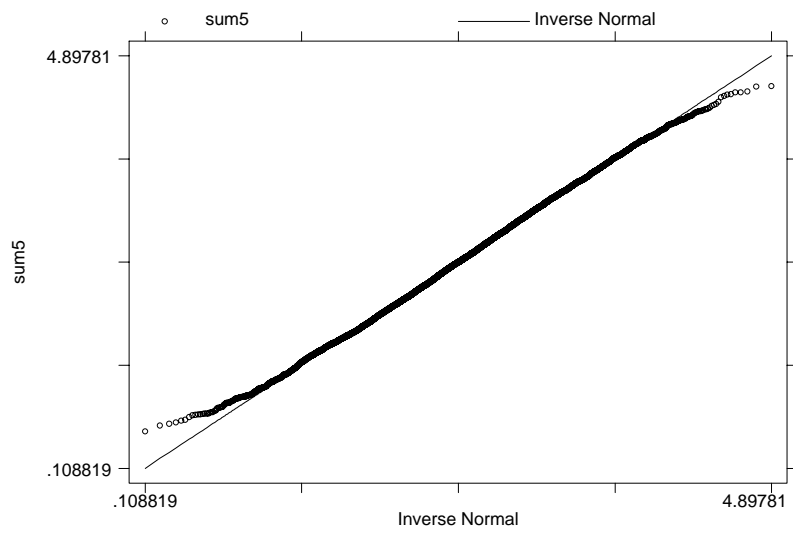
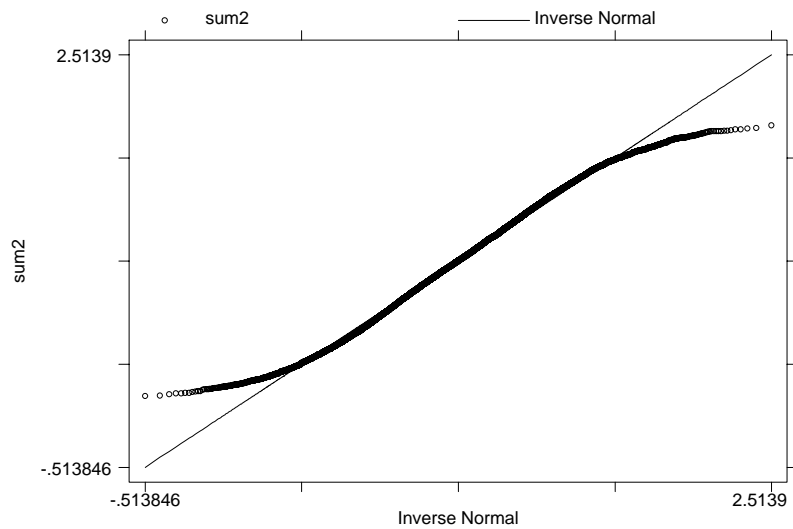
In the second plot, the random variable is the sum of 5 Uniform[0,1]

In the final Q-Q plot, the random variable is the sum of 10 Uniform[0,1]'s.

Note 2 things: first, the increasingly good approximation to the normal. Second, what is happening to the mean as the number of random variables added together to form the new one increases? (i.e. what is the expected value of a random variable that is the sum of 10 Uniform[0,1]'s? How could we prove this, using what tool that we have discussed?)

Note finally that this is the same result as in the discussion of the sample mean (which is often how the CLT is presented): Assume we compute the sample mean for several random variables. We then consider each of these sample means as values of a new random variable.

As the sample size increases, the distribution of this new variable converges on the normal. That is, the sampling distribution of the sample mean is normal.



More formally, let Y_1, Y_2, \dots, Y_n be any i.i.d. random variables with $E(Y) = \mu$ and $V(Y) = \sigma^2 < \infty$. If

$$U_n = \sqrt{n} \left(\frac{\bar{Y} - \mu}{\sigma} \right) \text{ where } \bar{Y} = \frac{1}{n} \sum_{i=1}^n Y_i$$

then the distribution function of U_n converges to a standard normal distribution function as $n \rightarrow \infty$. Another way of stating this is that \bar{Y} is *asymptotically normally distributed*:

$$\bar{Y} \overset{a}{\sim} N \left(\mu, \frac{\sigma^2}{n} \right)$$

Thus, if \bar{Y} is the mean of a random sample of size n from an infinite population with mean μ and std. dev. σ and n is large, then random variable:

$$Z = \frac{\bar{Y} - \mu}{\sigma_{\bar{Y}}} = \frac{\bar{Y} - \mu}{\sigma/\sqrt{n}}$$

Example (7.7): The achievement test scores of *all* high school seniors in New York have a mean of 60 and a variance of 64 (this is a sampling distribution). 100 students randomly drawn from one large high school have a mean score of 58. From the sampling distribution, find $P(\bar{Y} \leq 58)$. Does this suggest that this high school is inferior?

From the central limit theorem, we know that $\sqrt{n} \left(\frac{\bar{Y} - \mu}{\sigma} \right)$ is approximately a standard normal variable (Z). Thus:

$$P(\bar{Y} \leq 58) \approx P \left(Z \leq \frac{\sqrt{100}(58 - 60)}{\sqrt{64}} \right) = P(Z \leq -2.5) = .0062$$

from the table. This suggests that this particular sample has an average lower than the overall average of 60 (this logic will be formalized in the discussion of hypothesis testing in a few weeks).

Example (7.8): The service times for customers coming through a checkout counter in a store are distributed i.i.d. with mean 1.5 minutes and variance 1.0. What is the approximate probability that 100 customers can be served in 2 hours or less?

$$P \left(\sum_{i=1}^{100} Y_i \leq 120 \right) = P(\bar{Y} \leq 1.20)$$

Due to the large sample size (and about 100 is a good rule of thumb here), we can use the central limit theorem: $\bar{Y} \overset{a}{\sim} N \left(1.5, \frac{1.0}{\sqrt{100}} \right)$ (for the std. dev. remember that $\sigma_{\bar{Y}}^2 = \frac{\sigma^2}{n}$ and take the root). Thus:

$$P(\bar{Y} \leq 1.20) = P \left(\frac{\bar{Y} - 1.50}{1/\sqrt{100}} \leq \frac{1.20 - 1.50}{1/\sqrt{100}} \right) \approx P \left(Z \leq \sqrt{100}(1.2 - 1.5) \right) = P(Z \leq -3) = .0013$$

Homework problems: 7.1, 7.3, 7.4, 7.23, 7.24, 7.25, 7.30, 7.31, 7.33, 7.35, 7.37, 7.39

Sampling, Sampling Distributions and the Central Limit Theorem, Again

Suppose that the Bernoulli Widget Factory is making widgets, some of which are inevitably defective.

This is, of course, a Bernoulli system where each new widget is the outcome of a Bernoulli trial with some probability p of success (being defect-free) and probability $1-p$ of failure (i.e. being defective).

Think of this situation as if there is some hidden but real “Bernoulli Machine”, a data generating process (DGP) whose probability p governs the outcomes we observe in the real world.

Since the Bernoulli Machine is invisible, we don’t know what p is, but we’d like to find out. So, we take a random sample of n widgets and find that x of them are good.

Now, the proportion of successes in the sample should be somewhere around p and we call it \hat{p} , our estimate of p .

$$\hat{p} = \frac{x}{n}.$$

\hat{p} is simply the number of successes, x , in the sample divided by the sample size n . For example, maybe we found that 832 of the 1000 tacks we sampled were good giving us $\hat{p} = .832$.

We could ask: How good is this estimate?

But what does that mean?

We can’t know the precise difference between \hat{p} and p because we don’t know the value of p .

A better question is: if we took *many samples of 1000 widgets* and observed \hat{p} for each sample, how would those values of \hat{p} be distributed around p ?

Note that these \hat{p} values are looking more and more like a *random variable*: the selection of the n -unit sample is a random experiment and the observation \hat{p} is a numerical outcome.

If X is simply the number of successes in the sample, it is a binomial random variable and we define the observed proportion to be the random variable $\hat{P} = \frac{X}{n}$. (We use \hat{P} to indicate the random variable and \hat{p} for a particular sample.)

Knowing something about X , we can conclude a few facts about \hat{P} :

1. The mean of \hat{P} is $E(\hat{P}) = p$.
2. The standard deviation of \hat{P} is $\sigma_{\hat{P}} = \frac{\sqrt{p(1-p)}}{\sqrt{n}}$. (remembering the standard deviation for a binomial variable).
3. For large n , \hat{P} is approximately normal.

So, observed values of \hat{P} will be centered on p and their standard deviation is proportional to $\frac{1}{\sqrt{n}}$.

Since \hat{P} is approximately normal we can use our rules of thumb to conclude that approximately 68% of all estimates will fall within one standard deviation of the true value of p .

Going back to the widgets, with $n=1000$ and $p=0.85$, we get a standard deviation of:

$$\sigma_{\hat{P}} = \frac{\sqrt{.85(.15)}}{\sqrt{1000}} = 0.113.$$

So we expect about 68% of our estimates to fall in the interval: $0.8387 \leq \hat{P} \leq 0.8613$.

The standard deviation of \hat{P} is a measure of the *sampling error*. As we've seen, for the binomial \hat{P} , this sampling error is inversely proportional to \sqrt{n} . Increasing the sample size by a factor of 4 reduces the spread $\sigma(\hat{P})$ by a factor of 2.

Sample sizes for widgets, $p=0.85$.

n	1	4	16	25	100	10,000
\sqrt{n}	1	2	4	5	10	100
$\sigma(\hat{P})$.357	.1785	.089	.071	.0357	.0036

Already at $n=100$ you see that $\sigma_{\hat{P}}$ is down to 3.5%.

(Note the difference between an *estimate* and an *estimator*. An Estimator is a rule for getting estimates. So here, the estimator is the random variable $\hat{P} = \frac{X}{n}$).

Most of statistics involves the 4-step process above.

First, define the population with some unknown parameter.

Above this is p .

Second, find an estimator, its theoretical sampling distribution and standard deviation.

$$\text{Our estimator above is } \hat{P} \text{ with } E(\hat{P}) = p \text{ and } \sigma_{\hat{p}} = \frac{\sqrt{p(1-p)}}{\sqrt{n}}.$$

Third, actually draw a random sample and find the estimate.

Above, $\hat{p} = .832$.

Fourth, report the result and its sampling error.

Sampling Distribution of the Mean

Let's say we want to know the average length of time of a Senate committee meeting without having to time all of them.

So we select randomly n meetings and measure their lengths x_1, x_2, \dots, x_n .

If μ is the unknown length and σ is the standard deviation of the meeting length distribution then

$$E(X_i) = \mu \text{ and } \sigma(X_i) = \sigma.$$

Now we look at the sample mean, the average length of our selection of meetings.

$$\bar{X} = \frac{x_1 + x_2 + \dots + x_n}{n}.$$

We'd like to know how close this is to μ , meaning, if this sampling were done many times, what's the distribution of \bar{X} ?

$$\text{We know that } E(\bar{X}) = \mu \text{ and that } \sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}.$$

Do we know what the distribution of \bar{X} will look like? Do we need to know what the distribution of meetings looks like to answer that question?

Without knowing the distribution of meetings, we know that \bar{X} is distributed approximately normally.

This is the Central Limit Theorem.

If one takes random samples of size n from a population of mean μ and standard deviation σ , then, as n gets large, \bar{X} approaches the normal distribution with mean μ and standard deviation $\frac{\sigma}{\sqrt{n}}$.

$$\text{So, } P(a \leq \bar{X} \leq b) = P\left(\frac{a - \mu}{\sigma/\sqrt{n}} \leq Z \leq \frac{b - \mu}{\sigma/\sqrt{n}}\right).$$

What is important about this?

The CLT says that regardless of the shape of the original distribution – in this case meeting lengths – the taking of averages results in a normal distribution. To find the distribution of \bar{X} , we need know only the population mean and standard deviation.

There are two problems with the CLT, however.

First, it depends on a large sample size.

Second, to use it we need to know σ , the standard deviation of the population.

Of course, sample sizes are often small and σ usually unknown – for example, if we knew the standard deviation of all Senate Committee meetings we would probably also know the mean length and wouldn't bother with our sampling.

What we can then do is to estimate σ but taking the standard deviation of the *sample*

which is calculated as: $s = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2$.

Then, in place of the random variable $Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$ we substitute s for σ and define a new

random variable t by $t = \frac{\bar{X} - \mu}{s/\sqrt{n}}$.

Think of the random variable t as a substitute for the normal that allows us do the best we can given the sample size.

t – or “Student's t ” is more spread out than Z . It's flatter than the normal because the use of s introduces more uncertainty, making t “sloppier” than Z .

The amount of spread of the t depends on the sample size. The greater the sample size the more confident we can be that s is near σ , and the closer the t gets to the Z .

Just as there is a table of values of various areas under the Z curve, there are similar tables computing the area under t distributions of various shapes.