STAT 515 Lec 09

Sampling distributions and the central limit theorem

Karl Gregory

Sampling distributions

Remember that much of statistics has to do with learning about a population from a random sample. Here we consider the use of sample statistics as estimators of population parameters. Recall that sample statistics are values computed from a random sample and that population parameters are values belonging to the population *which we cannot know* unless we take a census.

We will focus on two situations in parallel:

1. Each draw from the population is a real number which we represent with the random variable X, and we are interested in the population parameters
$$\mu = \mathbb{E}X \quad \text{and} \quad \sigma^2 = \text{Var}(X),$$
which we call the *population mean* and the *population variance*, respectively.

2. Each draw from the population is a "success" or a "failure" which we encode in the random variable X as
$$X = \begin{cases} 1 & \text{if "success"} \\ 0 & \text{if "failure"}, \end{cases}$$
and we are interested in the population parameter
$$p = P(X = 1),$$
which we call the *population proportion*.

Case 1: Drawing real numbers

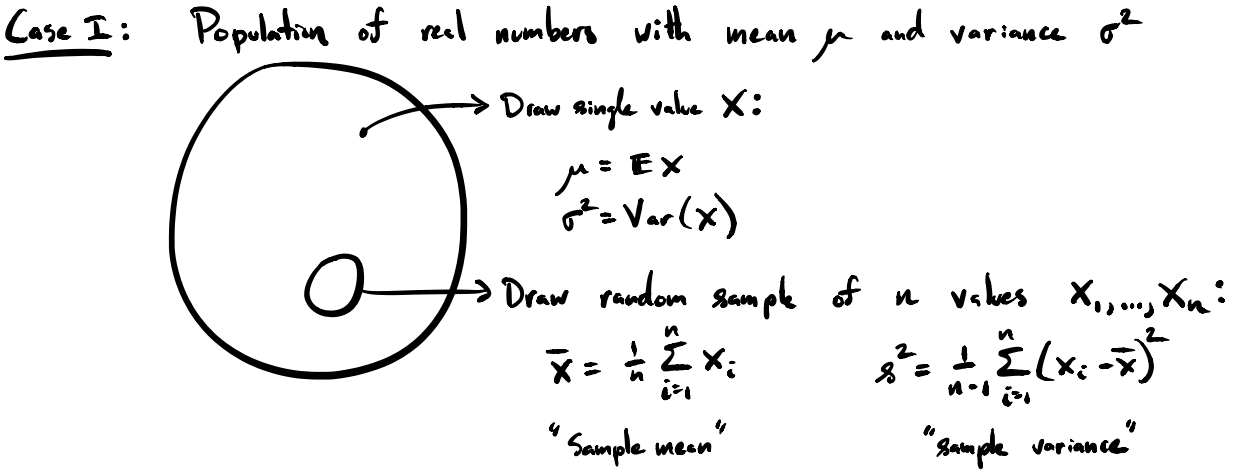
Consider drawing a random sample of size n from the population and let the n real numbers be represented by the random variables
$$X_1, \ldots, X_n.$$

Now define the sample quantities

$$\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i \quad \text{and} \quad S_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2,$$

which we call the *sample mean* and the *sample variance*, respectively.

The sample statistics $\bar{X}$ and s^2 are estimators of the population parameters μ and σ^2; that is, $\bar{X}$ and s^2 are guesses, from the data in the random sample, of the unknown values of μ and σ^2 (We will find that they are good guesses).



Case 2: Drawing successes and failures

Consider drawing a random sample of size n from the population and let the n "successes" and "failures" in the sample be encoded as ones and zeroes in the random variables

$$X_1, \ldots, X_n$$

such that

$$X_i = \begin{cases} 1 & \text{if draw } i \text{ is a "success"} \\ 0 & \text{if draw } i \text{ is a "failure"} \end{cases}$$

for $i = 1, \ldots, n$.

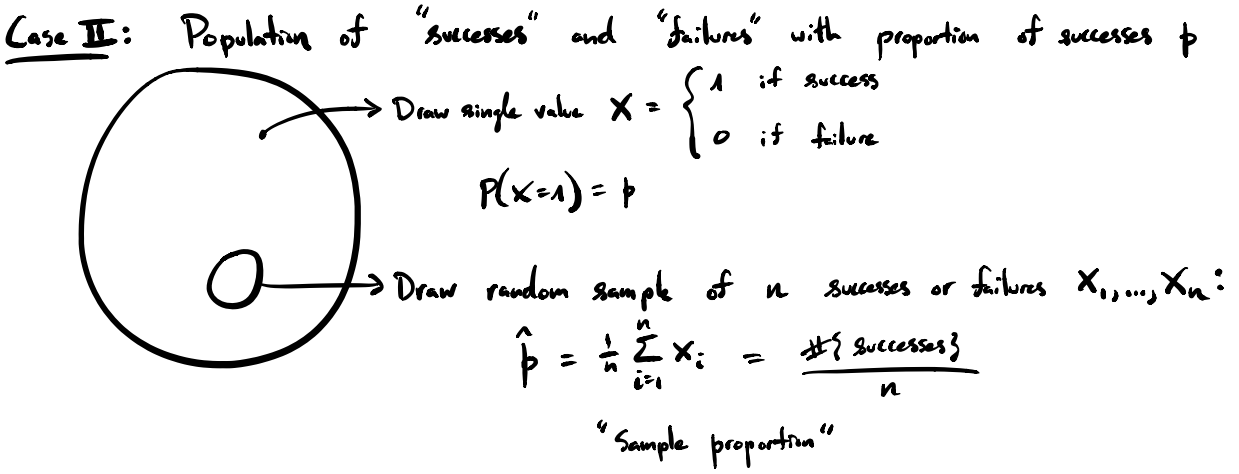
Now define the sample quantity

$$\hat{p} = \frac{1}{n}\sum_{i=1}^{n} X_i = \frac{\#\{\text{successes}\}}{n},$$

which we call the *sample proportion*.

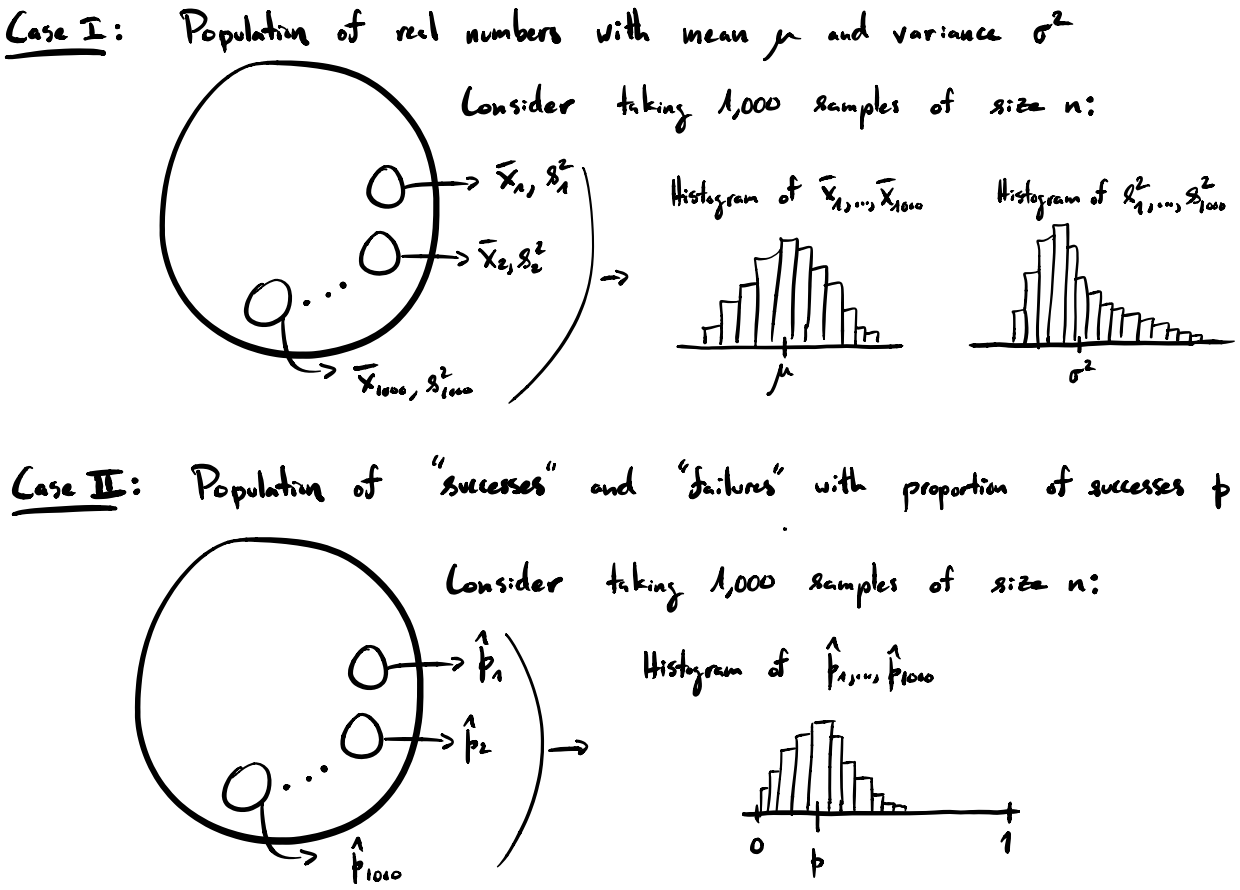
The sample statistic $\hat{p}$ is an estimator of the population parameter p; that is, $\hat{p}$ is a guess, from the data in the random sample, of the unknown value of p (We will find that it is a good guess).

Note that if another random sample were drawn, it is likely that a different value of $\hat{p}$ would be obtained.



Learning about the population from the random sample

Given that the values of $\bar{X}$ and s^2 or $\hat{p}$ change from sample-to-sample, how useful can they be to inform us about their population counterparts μ and σ^2 or p? To answer this question we must do some thought experiments: We must ask ourselves what values of $\bar{X}$ and s^2 or $\hat{p}$ we would get if we took a very large number of random samples—that is, if we repeated our random sampling over and over again. How often would certain values of $\bar{X}$ and s^2 or $\hat{p}$ occur or how often would they fall within certain ranges? Would they be close to their population counterparts, and if so, how close can we expect them to be? If we made histograms of their values over many repeated samples, what would those histograms look like? Put in a technical way, what *probability distributions* would these quantities have if we were to repeat our sampling many many times? These distributions, whatever they may look like, are called *sampling distributions*, as they reflect how sample statistics vary from sample to sample.



One goal of considering these *sampling distributions* is to be able to make statements like

1. "We are pretty sure† μ is something like $\bar{X}$ plus or minus something‡."
2. "We are pretty sure† p is something like $\hat{p}$ plus or minus something‡."

†: Studying the sampling distribution lets us make "pretty sure" more precise.
‡: Studying the sampling distribution enables us to find an appropriate "something", which we will later call a *margin of error*.

Sampling distribution of the mean

To understand what the sample mean $\bar{X}$ can tell us about the population mean μ, we must understand how $\bar{X}$ behaves from sample to sample. Its behavior is described by its *sampling distribution*.

What we must imagine is repeating our statistical experiment many times—drawing many random samples of size n and computing the mean of each one. If we were to build a histogram of these sample means, what would it look like?

Expected value and variance of the sample mean

Two things are always true of the sample mean:

Sampling distribution result: Mean and variance of the sample mean

Let $X_1, \ldots, X_n$ be a random sample from a population with mean μ and variance σ^2 and let $\bar{X} = (X_1 + \cdots + X_n)/n$. Then

$$\mathbb{E}\bar{X} = \mu \quad \text{and} \quad \operatorname{Var}\bar{X} = \frac{\sigma^2}{n}.$$

What will happen as we take a larger and larger sample size? The variance of the sample mean will become very small. This tells us that if we take very large samples, our sample means will not change very much from sample to sample. In other words, we *expect* the sample mean from a larger sample to be closer to the population mean μ than the sample mean from a smaller sample.

The mean of a sample from a Normal population

First suppose that we draw our sample of size n from a population of values which follow a Normal distribution with mean μ and variance σ^2.

Sampling distribution result: Mean of random sample from a Normal population

Let $X_1, \ldots, X_n$ be a random sample from a population which has the Normal distribution with mean μ and variance σ^2 and let $\bar{X} = (X_1 + \cdots + X_n)/n$. Then

$$\bar{X} \sim \text{Normal}\left(\mu, \frac{\sigma^2}{n}\right).$$

How might we use this result? If $\bar{X}$ has the Normal distribution with mean μ and variance σ^2/n, we may find $P(a \leq X \leq b)$ as follows:

1. Transform a and b from the $\bar{X}$ world to the Z world (the number-of-standard-deviations

world) by

$$a \mapsto \frac{a-\mu}{\sigma/\sqrt{n}} \quad \text{and} \quad b \mapsto \frac{b-\mu}{\sigma/\sqrt{n}},$$

since

$$\begin{aligned} P\left(a \le \bar{X} \le b\right) &= P\left(\frac{a-\mu}{\sigma/\sqrt{n}} \le \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \le \frac{b-\mu}{\sigma/\sqrt{n}}\right) \\ &= P\left(\frac{a-\mu}{\sigma/\sqrt{n}} \le Z \le \frac{b-\mu}{\sigma/\sqrt{n}}\right). \end{aligned}$$

Note that if $a = \pm\infty$ then $(a-\mu)/(\sigma/\sqrt{n}) = \pm\infty$ and if $b = \pm\infty$ then $(b-\mu)/(\sigma/\sqrt{n}) = \pm\infty$.

2. Look up probabilities for Z on the z-table.

Exercise. Let X be the number of minutes of cell-phone talk time in the last month of a randomly selected USC undergraduate. Suppose that the cell phone talk times of USC undergraduates are Normally distributed with mean $\mu = 300$ and variance $\sigma^2 = 50^2$.

1. Find $P(|X - 300| > 50)$.

   Answer: First, let's figure out what $|X - 300| > 50$ means: it means that X is more than 50 away from the mean 300, so we want $P(X < 250 \text{ or } X > 350)$. For the values 250 and 350 we compute Z, the number of standard deviations from the mean:

   $$Z = \frac{X-\mu}{\sigma} \quad \text{gives} \quad \frac{250-300}{50} = -1 \quad \text{and} \quad \frac{350-300}{50} = 1.$$

   We get from the Z table that $P(Z < -1) = (Z > 1) = 0.1587$, so the answer is $2(0.1587) = 0.3174$.

2. Suppose you sampled 4 USC undergraduates and took the average of their cell phone talk times. Let the average of the four talk times be $\bar{X}$. What is $P(|\bar{X} - 300| > 50)$?

   Answer: Like before, we interpret this as $P(\bar{X} < 250 \text{ or } \bar{X} > 350)$. We again compute the number of standard deviations of 250 and 350 from the mean, but this time the standard deviation of our random variable is $\sigma/\sqrt{n}$. The variance of $\bar{X}$ is σ^2/n, so our Z is going to be different:

   $$Z = \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \quad \text{gives} \quad \frac{250-300}{50/2} = -2 \quad \text{and} \quad \frac{350-300}{50/2} = 2.$$

   We get from the Z table that $P(Z < -2) = P(Z > 2) = 0.0228$, so the answer is $2(0.0228) = 0.0456$.

3. Sample 9 USC undergraduates and let $\bar{X}$ be the mean of their cell phone talk times. What is $P(|\bar{X} - 300| > 50)$?

   Answer: With a sample size of $n = 9$, the standard deviation of the mean $\bar{X}$ is smaller—its variability around the true mean decreases as the sample size grows. Now

   $$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \quad \text{gives} \quad \frac{250 - 300}{50/3} = -3 \quad \text{and} \quad \frac{350 - 300}{50/3} = 3.$$

   We get from the Z table that $P(Z < -3) = P(Z > 3) = 0.0013$, so the answer is $2(0.0013) = 0.0026$.

What is happening to the distribution of $\bar{X}$ as the sample size increases? It is getting narrower, concentrating around the mean. For larger samples, the chances are smaller that we will get a value of $\bar{X}$ which is far from the mean.

The most important theorem in statistics

Probably the most important theorem in statistics is the central limit theorem. This theorem tells us that the sample mean behaves like a Normal random variable if the sample size is large enough—even if the population itself is not Normal!

Sampling distribution result: Central limit theorem

Let $X_1, \ldots, X_n$ be a random sample from a population with mean μ and variance $\sigma^2 < \infty$ and let $\bar{X} = (X_1 + \cdots + X_n)/n$. Then

$$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \text{ has a distribution more and more like the } \text{Normal}(0, 1)$$

distribution for larger and larger sample sizes n.

Remark 1. *How large should the sample size n be before we can invoke the central limit theorem? A rule of thumb is that we can safely treat $(\bar{X} - \mu)/(\sigma/\sqrt{n})$ like a* Normal$(0, 1)$ *random variable if $n \geq 30$.*

Exercise. Let X be the marathon time of a randomly selected runner of the next Columbia marathon. The distribution is skewed to the right and has mean 4.5 hours and standard deviation 2 hours. Suppose you take a random sample of 30 finishers. What is the probability that the mean of the finishing times of the 30 runners is less than 4.25 hours?

Answer: Even though the marathon times are not Normally distributed, the Z-score $(\bar{X} - \mu)/(\sigma/\sqrt{n})$ of the sample mean should behave like a Normal$(0, 1)$ random variable. Thus we

can get $P(\bar{X} < 4.25)$ using the Normal distribution:

$$Z = \frac{4.25 - 4.5}{\sqrt{4/30}} = -0.68,$$

and $P(Z < -0.68) = 0.2483$. So the answer is

$$P(\bar{X} < 4.25) = 0.2483.$$

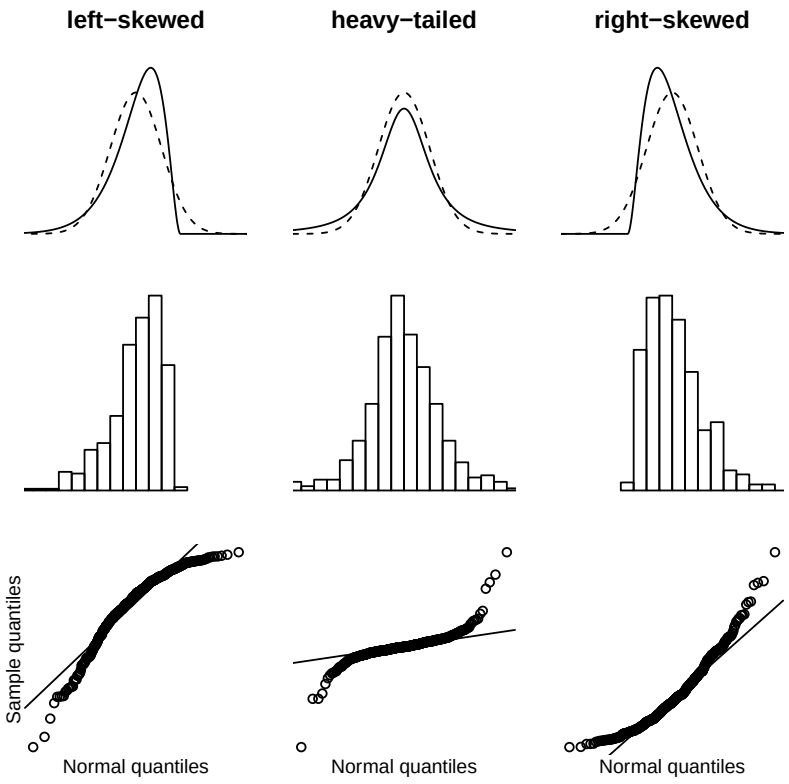
Example. Let X be the number of ships which come through a set of locks in an afternoon, and the mean and standard deviation of X are 6 and 1, respectively. Suppose you observe the locks on five randomly selected afternoons and compute $\bar{X}$, the mean of the numbers of ships you counted on the 5 afternoons. What is $P(\bar{X} > 7)$?

Answer: We cannot compute it, because the sample size n is small ($5 < 30$), and the central limit theorem holds only for large sample sizes.

More ways to describe probability distributions

The Normal distribution has a bell-shaped probability density function. We often describe distributions by the way their probability density functions differ in shape from that of the Normal distribution: A *left-skewed* distribution produces more observations to the far left of the mean than the Normal distribution, a *heavy-tailed* distribution produces more extreme values far away from the mean in both directions, a *right-skewed* distribution produces more observations to the far right of the mean than the Normal distribution.

The plots below show probability density functions for a left-skewed, heavy-tailed, and a right-skewed distribution (solid lines) with the Normal probability density function (dashed line) overlaid. Below these plots are histograms from a sample of size $n = 500$ drawn from the respective distributions. In the bottom row of the figure, Normal QQ plots are given comparing the quantiles of the sample to the quantiles of the Normal distribution.



The central limit theorem says that even when the population has a distribution which is left-skewed, heavy-tailed, right-skewed, or even which differs from the Normal distribution in some other way, the mean of a large enough sample may be treated as a Normal random variable. This is taken advantage of all the time in statistical practice.

Sampling distribution of the sample proportion

We can get the expected value and variance of the sample proportion by expressing it as a mean of random variables which are equal to zero or one. Moreover, since can express the sample proportion $\hat{p}$ as a mean we can use the central limit theorem to treat it as a random variable having a Normal distribution when the sample size is large.

Expected value and variance of the sample proportion

Suppose, as before, that we encode the outcome of a Bernoulli trial in the random variable X such that

$$X = \begin{cases} 1 & \text{if outcome a "success"} \\ 0 & \text{if outcome a "failure".} \end{cases}$$

If the Bernoulli trial has success probability p, then we have $P(X = 1) = p$ and $P(X = 0) = 1 - p$. We can compute

$$\mathbb{E}X = p \text{ and } \text{Var}(X) = p(1 - p).$$

Suppose we ran the Bernoulli trial n times independently and got $X_1, \ldots, X_n$. Then the sample proportion $\hat{p}$ of successes can be defined as

$$\hat{p} = \frac{1}{n} \sum_{i=1}^{n} X_i = \frac{\#\{\text{successes}\}}{n}.$$

Given that $\hat{p}$ is simply the sample mean of the Bernoulli random variables $X_1, \ldots, X_n$, its expected value is p and its variance is $p(1 - p)/n$. We express this in the following result:

Sampling distribution result: Mean and variance of a sample proportion

Let $X_1, \ldots, X_n$ be independent random variables equal to 1 with probability p and equal to 0 with probability $1 - p$ and let $\hat{p} = (X_1 + \cdots + X_n)/n$. Then

$$\mathbb{E}\hat{p} = p \quad \text{and} \quad \text{Var}(\hat{p}) = \frac{p(1 - p)}{n}.$$

Central limit theorem for the sample proportion

Additionally, we can apply the central limit theorem to $\hat{p}$ which lets us treat $\hat{p}$ like a Normally distributed random variable when the sample size n is large enough.

Sampling distribution result: Central limit theorem for sample proportion

Let $X_1, \ldots, X_n$ be independent random variables equal to 1 with probability p and equal to 0 with probability $1 - p$ and let $\hat{p} = (X_1 + \cdots + X_n)/n$. Then

$$\frac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{n}}} \text{ has a distribution more and more like the } \text{Normal}(0, 1)$$

distribution for larger and larger sample sizes n.

This result tells us that if n is large, we have

$$\hat{p} \overset{\text{approx}}{\sim} \text{Normal}\left(p, \frac{p(1-p)}{n}\right),$$

so we can compute approximations to probabilities of the form $P(a \leq \hat{p} \leq b)$ as follows:

1. Transform a and b from the $\hat{p}$ world to the Z world (the number-of-standard-deviations world) by

$$a \mapsto \frac{a-p}{\sqrt{p(1-p)/n}} \quad \text{and} \quad b \mapsto \frac{b-p}{\sqrt{p(1-p)/n}},$$

since

$$\begin{aligned} P\left(a \leq \hat{p} \leq b\right) &= P\left(\frac{a-p}{\sqrt{p(1-p)/n}} \leq \frac{\hat{p}-p}{\sqrt{p(1-p)/n}} \leq \frac{b-p}{\sqrt{p(1-p)/n}}\right) \\ &= P\left(\frac{a-p}{\sqrt{p(1-p)/n}} \leq Z \leq \frac{b-p}{\sqrt{p(1-p)/n}}\right). \end{aligned}$$

Note that if $a = \pm\infty$ then $\frac{a-p}{\sqrt{p(1-p)/n}} = \pm\infty$ and if $b = \pm\infty$ then $\frac{b-p}{\sqrt{p(1-p)/n}} = \pm\infty$.

2. Look up probabilities for Z on the z-table.

Remark 2. *How large should n be before we can invoke the central limit theorem for the sample proportion $\hat{p}$? A rule of thumb is that we can safely treat $(\hat{p}-p)/(\sqrt{p(1-p)/n})$ as a* Normal$(0,1)$ *random variable if*

$$np \geq 5 \quad \textit{and} \quad n(1-p) \geq 5.$$

Example. Suppose you take a random sample of 15 USC undergraduates and you ask each one if they are registered to vote. Let $\hat{p}$ be the proportion in your sample who are registered to vote. Supposing that the true proportion of USC undergraduates who are registered to vote is 0.6, What is the probability that $\hat{p}$ of your sample is greater than 0.7?

Answer: For a sample of size $n = 15$ and with $p = 0.6$, we have

$$15(0.6) = 9 \quad \text{and} \quad 15(0.4) = 6,$$

so according to the rule of thumb, $(\hat{p}-p)/\sqrt{p(1-p)/n}$ should behave approximately like a Normal$(0,1)$ random variable. Now

$$Z = \frac{\hat{p}-p}{\sqrt{\frac{p(1-p)}{n}}} \quad \text{gives} \quad \frac{0.7-0.6}{\sqrt{\frac{0.6(1-0.6)}{15}}} = 0.79.$$

We get from the table that $P(Z > 0.79) = 0.2148$. So the answer is

$$P(\hat{p} > 0.7) \approx 0.2148.$$

Another answer: We could also use the Binomial distribution to get the exact answer. The event $\hat{p} > 0.7$ corresponds to observing 11 or more successes out of the 15 Bernoulli trials. So if Y is the number of successes, $P(\hat{p} > 0.7) = P(Y \geq 11) = 1 - P(Y < 10)$. We can compute $P(X < 10)$ in R using the command

```
pbinom(q=10,size=15,prob=.6)
```

We get $P(Y < 10) = 0.7827$, so the answer is

$$P(\hat{p} > 0.7) = P(Y \geq 11) = 1 - 0.7827 = 0.2173.$$

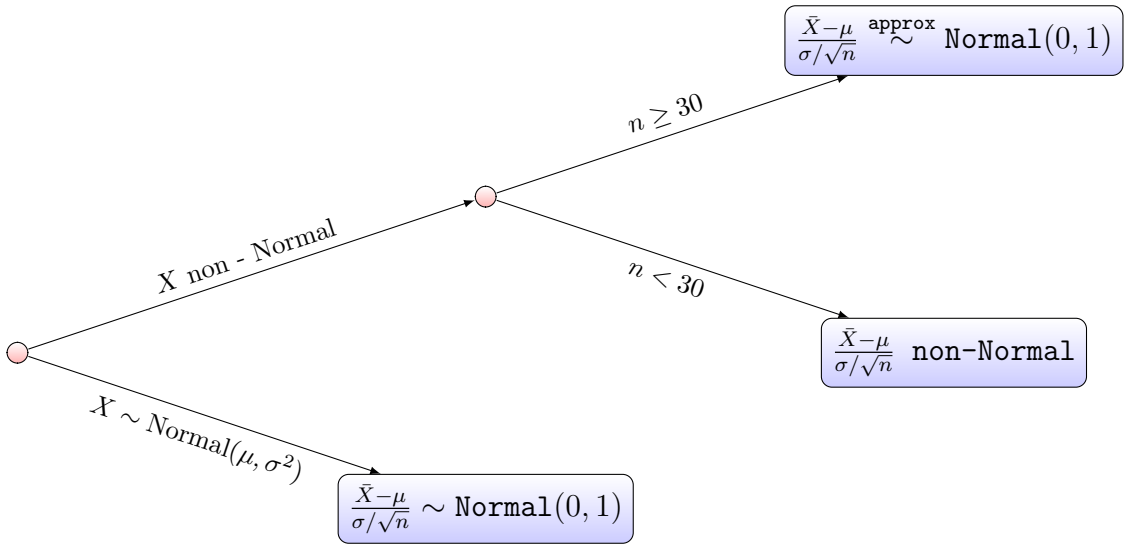
It is close to the first answer, which came from using the central limit theorem.

Sampling distribution summaries for $\bar{X}$ and $\hat{p}$

The sampling distribution of the standardized sample mean

$$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$$

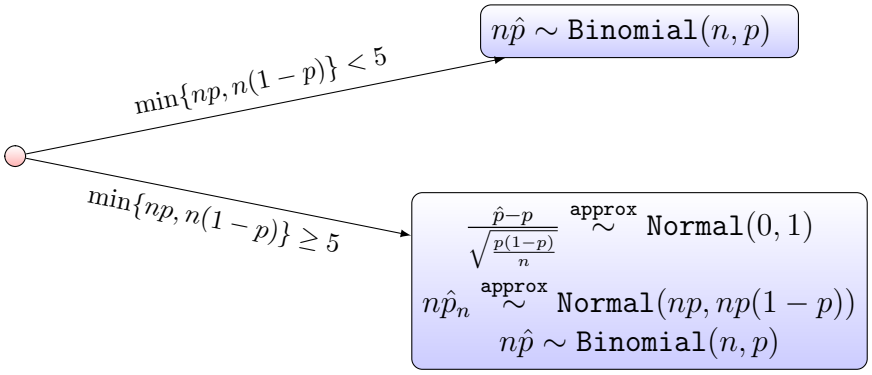
is summarized in this diagram:



The sampling distribution of the standardized sample proportion

$$\frac{\hat{p} - p}{\sqrt{p(1-p)/n}}$$

is summarized in this diagram:



Recall that $n\hat{p} = Y$, the number of successes in n Bernoulli trials, so saying that $n\hat{p} \sim$ Binomial(n, p) is nothing new, and it is always true, no matter what n is.

Further examples of the central limit theorem

Exercise. Suppose X is the time between phone calls to a customer service call center every hour, and suppose it follows the exponential distribution with mean equal to $1/20$. Suppose we observe the next 30 time intervals between calls and let $\bar{X}$ be the mean length of the 30 time intervals. What is $P(\bar{X} > .075)$?

Answer: For the Exponential(λ) distribution, we have $\mu = 1/\lambda$ and $\sigma^2 = 1/\lambda^2$. According to the central limit theorem, $(\bar{X} - \lambda^{-1})/(\lambda^{-1}/\sqrt{n})$ should behave approximately like a Normal$(0, 1)$ random variable. So, with $\lambda = 20$, we get

$$Z = \frac{\bar{X} - \lambda^{-1}}{\lambda^{-1}/\sqrt{n}} = \frac{0.075 - 0.05}{\sqrt{0.0025/30}} = 2.74,$$

and $P(Z > 2.74) = 0.0031$.

Exercise. Suppose X is the number of phone calls to a call center every hour, and suppose it follows the Poisson distribution with $\lambda = 20$. Suppose we observe the call center during 25 randomly selected hours and let $\bar{X}$ the mean number of calls per hour over the 25 hours. What is $P(\bar{X} < 18)$?

Answer: For the Poisson distribution, we have $\mu = \lambda$ and $\sigma^2 = \lambda$. According to the central limit theorem, $(\bar{X} - \lambda)/(\sqrt{\lambda/n})$ should behave approximately like a Normal$(0, 1)$ random variable. So, with $\lambda = 20$, we get

$$Z = \frac{\bar{X} - \lambda}{\sqrt{\lambda/n}} = \frac{18 - 20}{\sqrt{20/25}} = -2.24,$$

and $P(Z < -2.24) = 0.0125$.