# Sampling Distributions

## What Is a Statistic?

- It is a function of observable random variables.
- Random sampling is used to observe these random variables.
- The purpose of a statistic is to estimate a population parameter.

  Examples:

  – $\bar{X}$ for $\mu$

  – $s^2$ for $\sigma^2$



## Simple Random Sampling

- Suppose a sample of size $n$ has to be chosen from a population of size $N$.
- There are $\binom{N}{n}$ possible samples, since they are chosen without replacement.
- Simple random sampling is a procedure where each subset of size $n$ from the population is equally likely to be chosen as a sample.
- This is one of the most common methods of sampling (although it is not always possible).
- When we talk about "random sampling" in this course, we will generally be referring to "simple random sampling."



- Simple random sampling is analogous to drawing a set of cards from a box of cards without replacement where the box of cards is analogous to the population and the set of cards drawn is analogous to a sample.
- If the population size $N$ is very large compared to the sample size $n$, then drawing without replacement and drawing with replacement are almost equivalent.
- In most of the cases $N$ is very large relative to $n$. So in most cases we can think of a sample as a realization of a set of independent and identically distributed random variables.



## Sampling Distributions

- We're interested in the probability distributions for statistics that we use. These distributions are also called sampling distributions.
- Understanding the distribution for the statistic is critical for understanding how good your estimate (obtained using the statistic) is.
- Remember that statistics are random variables, because we're dealing with random samples (and the sample determines the value of the statistic).

We would like to have information about different aspects like mean, variance, shape *etc.* of the distribution for a statistic.

**Mean and Variance of $\bar{X}$**

Suppose we have a random sample of size $n$ from a population with mean $\mu$ and variance $\sigma^2$. Show that the expected value of $\bar{X}$ is $\mu$ and that the variance of $\bar{X}$ is $\frac{\sigma^2}{n}$. Note that this is true for any population as long as the mean and variance of the population exist.



**Sampling Dist. for $\bar{X}$, Normal Pop., Known $\sigma$**

Since we know the mean and variance of $\bar{X}$ for any population, we know those in particular for normally distributed population. Now we would like to know about other aspects of the distribution. Does $\bar{X}$ have a distribution that we are already familiar with (like the normal)? If not, what can we say about it?

- Any linear combination of normal random variables is normally distributed (Ref thm. 6.3, pp. 305-6).
- Since the sample mean $\bar{X}$ is a linear combination of normal random variables, $\bar{X}$ is normally distributed.
- We conclude that $\bar{X} \sim N(\mu, \frac{\sigma^2}{n})$.



**Example: Gasoline Prices**

Assume the prices for gasoline are normally distributed with standard deviation \$0.10. Suppose that a random sample of 25 gasoline stations is selected.

What is the probability that the simple random sample will provide a sample mean within 3 cents, \$0.03, of the population mean?



**Sampling Distribution for Sum of Squares for Normal Variables**

If $X_1, X_2, \ldots, X_n$ is a random sample from $N(\mu, \sigma^2)$ population, then

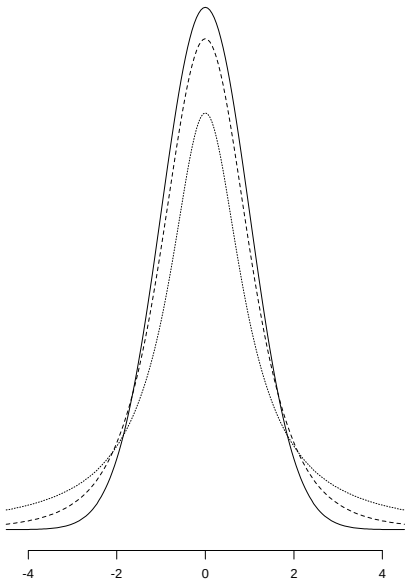
- $\frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i-\mu)^2$ has a chi-square distribution with $n$ degrees of freedom (sometimes written as $\chi^2_n$). Recall that chi-square is a particular type of gamma distribution.
- $(n-1)\frac{s^2}{\sigma^2} = \frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i-\bar{X})^2$ has a chi-square distribution with $n-1$ degrees of freedom. Here $s$ is the sample SD.
- Also $\bar{X}$ and $\frac{s^2}{\sigma^2}$ are independent.

## Student's $t$ Distribution

- If $Z \sim N(0,1)$ and $Y \sim \chi^2_n$, and also $Z$ and $Y$ are independent, then $\frac{Z}{\sqrt{Y/n}}$ has a $t$ distribution with $n$ degrees of freedom (d.f.)
- The $t$ distribution is also symmetric and bell-shaped like the normal distribution, but the $t$ distribution has more area in the tails and shorter at the center.
- As $n \to \infty$, the $t$ distribution with $n-1$ d.f. converges to the normal.
- Standardized table for $t$ includes d.f. as rows and gives values for certain key quantiles in columns.
- Standard $t$ distribution with $n$ degrees of freedom has mean 0 and variance $\frac{n}{n-2} > 1$.



## How Does the $t$ Distribution Look?






## Examples Using the t Table

- $P(1.476 < T^5 < 2.015)$
- $P(T^8 < -2.5)$
- What value marks the 99th percentile of the t distribution with 10 d.f.?



## Sampling Dist. for $\bar{X}$, Normal Pop., Unknown $\sigma$

- Previously, we assumed that we knew the population standard deviation $\sigma$. However, the more common case is that $\sigma$ is unknown. What is the distribution of $\bar{X}$ in that case?
- It is still $N(\mu, \frac{\sigma^2}{n})$, but some problems in inference about $\mu$ require the distribution of $\bar{X}$ to be completely known except the value of $\mu$. In those cases knowing $\bar{X} \sim N(\mu, \frac{\sigma^2}{n})$ is useless because we still do not know $\sigma$.

- To bypass this problem, we employ our knowledge about the distribution of the sample standard deviation $s$. We estimate $\sigma$ by $s$. Then by using the fact that $\bar{X} \sim N(\mu, \frac{\sigma^2}{n})$ and $(n-1)\frac{s^2}{\sigma^2} \sim \chi^2_{n-1}$, and $\bar{X}$ and $s$ are independent, we see that $\frac{\bar{X}-\mu}{s/\sqrt{n}}$ has a $t$ distribution with $n-1$ degrees of freedom. Again recall the formula for $s$.

$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$



**Example: Gasoline Prices (cont.)**

Now assume the prices for gasoline are normally distributed with *unknown* standard deviation. Suppose that a random sample of 25 gasoline stations is selected; the sample standard deviation is \$0.10.

What is the probability that the simple random sample will provide a sample mean within 3 cents, \$0.03, of the population mean?



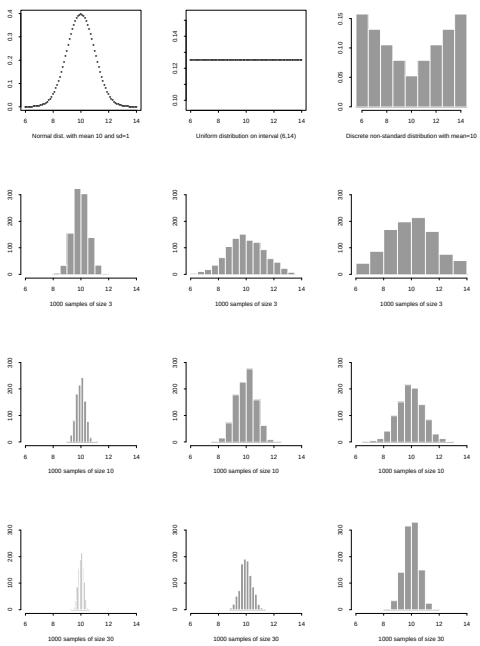
**Central Limit Theorem (CLT)**

What if the population from which we are drawing the sample doesn't have a normal distribution?

- Often, we have no idea how the population is distributed.
- Central Limit Theorem (CLT) tells us that as $n \to \infty$, the distribution of $\bar{X}$ converges to $N(\mu, \frac{\sigma^2}{n})$.
- The closer to normally distributed the population is, the smaller the size that $n$ needs to be before we say that $\bar{X}$ has a normal distribution.
- In most cases sample size $n \geq 30$ is big enough.



**Visual: Central Limit Theorem**



**Example Using the CLT**

Standardized test scores for a certain test are assumed to have variance 100 and mean 100. What is the probability that a sample of 36 standardized scores will yield a sample mean of 105 or more?



**Another Example Using the CLT**

A freight elevator lifts 520 pounds. The average package is 10 pounds; standard deviation is 5 pounds. What is the probability the elevator will be overloaded if 49 packages are placed in the elevator?



**Sampling from a Population of "0"s and "1"s**

Consider a survey in which you want to ask one "yes" or "no" question. Examples: "Do you own your home?", " Are you more that 20 years old?", etc. You are interested in the proportion of people responding "yes".

- Envision the population as a box of tickets, each marked with either "0" or "1".
- A member of the population who can answer "yes" to the question of interest is like a "1" ticket.
- Taking a sample of size $n$ from this population is like taking $n$ tickets from the box.
- Finding the number who answered "yes" is same as adding up the values on the $n$ tickets.



**Connecting the Binomial to the Survey Problem**

- The $n$ sampled tickets can be viewed as $n$ trials (practically independent if $n$ is much smaller than the total population $N$).
- Then the number of "1" in the sample is the number of successes. The percentage of "1" tickets in the box is the probability of success for the trials.
- Binomial p.m.f. can be used to calculate the probability of observing a given number of "yes" answers in a sample from the population.
- If the sample size is large, it becomes difficult to calculate $\binom{n}{x}p^x(1-p)^{n-x}$ by using a calculator.

## Can We Use the CLT?

- The sample proportion is actually the sample mean (remember: "yes" answers are "1"s, "no" answers are "0"s).
- If the number of successes in the sample is denoted as $Y$, the sample proportion is $\hat{p} = \frac{Y}{n}$.
- This means CLT can be used if the sample size is large enough.
- General rule for determining how large $n$ should be depends on $p$; both endpoints of $p \pm 2\sqrt{\frac{pq}{n}}$ should be between 0 and 1.
- In order to use the CLT, we need to know the expected value and variance for $\hat{p}$.



## Mean and Variance of Sample Proportion

What are the mean and variance for $\hat{p} = \frac{Y}{n}$?



## Visual: Sampling Distribution for Sample Proportion

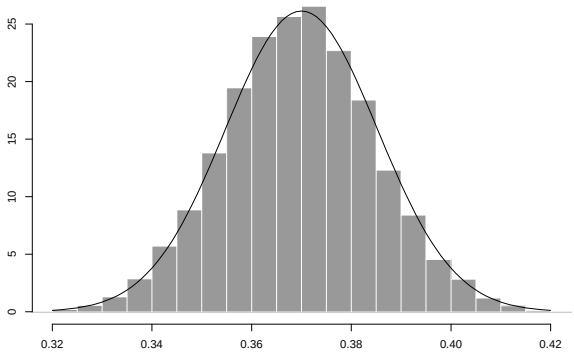



Figure 1: Histogram of 10000 sample proportions



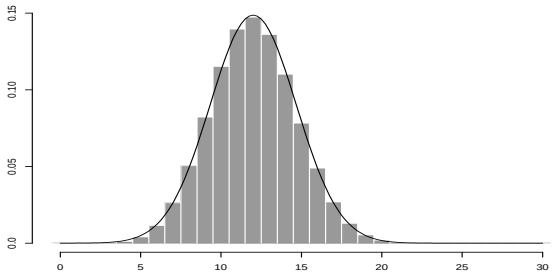
## Basics of Using CLT for Proportions

Assume women constitute 37% of all union members, so that $p = 0.37$. A simple random sample of 1000 union members is selected.

What is the probability that the sample proportion of women will be within $\pm 0.03$ of the population proportion?

## Normal Approximation to Binomial

What's the probability that a sample of 30 voters contains 15 or more who voted Democratic, given that 40% of the population voted Democratic? With patience, could solve this by calculator. Can also approximate using the normal distribution.




Why is the number of successes $Y$ normally distributed? $Y$ is a linear function of a normally distributed random variable $\hat{p} = \frac{Y}{n}$, so $Y$ must be normally distributed.



## The Continuity Correction

The normal approximation of binomial distribution can be improved by proper choice of boundary points. Note that the bars in the probability histogram for the binomial distribution are centered at the integer values and range $\pm.5$ around those values. For example, the bar over 0 is centered at 0 and ranges from $-.5$ to .5. The probability of a particular value is the area of the bar that is centered at that value. So, when we find the probability that a binomial variable is $k$, we find the area of the bar over the interval $(k - .5, k + .5)$. This area is approximated by the area over that interval under the corresponding normal curve.



## Some Examples

- In the discrete case, phrases like "exactly 25" can be rewritten as $P(24.5 < Y < 25.5)$ to make the transition to the continuous case.
- Phrases like "less than 21" are equivalent to their counterparts of the form "20 or less." Then look for $P(Y < 20.5)$.
- To find the probability of "more than 14" (or equivalently "15 or more"), find $P(Y > 14.5)$.



## Example: Using the Continuity Correction

A random sample of 100 children is taken from the children born this year. What is the chance there will be exactly 57 boys? Find the exact answer, then make an approximation using the normal. (Assume male and female children are equally likely.)

**Example: Using the Continuity Correction**

What's the probability that a sample of 30 voters contains 15 or more who voted Democratic, given that 40% of the population voted Democratic?



**Summary of Sampling Distributions**

Suppose that $X_1, X_2, \ldots, X_n$ is a random sample from a population with mean $\mu$ and variance $\sigma^2$. The sample mean is $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$. The sample standard deviation is $s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2}$.

**1.** $E(\bar{X}) = \mu$ and $V(\bar{X}) = \frac{\sigma^2}{n}$

**2.** If the population is normal,

- $\bar{X} \sim N(\mu, \frac{\sigma^2}{n})$. (Use this for known $\sigma$).
- $\frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i - \mu)^2 \sim \chi^2_n$
- $(n-1)\frac{s^2}{\sigma^2} = \frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i - \bar{X})^2 \sim \chi^2_{n-1}$.
- $\bar{X}$ and $\frac{s^2}{\sigma^2}$ are independent.
- $\frac{\bar{X}-\mu}{s/\sqrt{n}} \sim t_{n-1}$. (Use this for unknown $\sigma$).



**3.** If the population is non-normal with mean $\mu$ and known standard deviation $\sigma$, using CLT:

- If $n \geq 30$, $\bar{X}$ is approximately $N(\mu, \frac{\sigma^2}{n})$.
- When you have no information about the shape of the population distribution, you shouldn't assume you can use the CLT unless $n \geq 30$.

**4.** If the population consists of all "1"s and "0"s and $n$ is large enough so that both endpoints of $p \pm \sqrt{\frac{p(1-p)}{n}}$ fall in the interval (0,1) (using CLT):

- The sample proportion $\hat{p}$ is normally distributed with mean $p$ (population proportion) and standard deviation $\sqrt{\frac{p(1-p)}{n}}$.
- The number of "1"s, $Y$, is normally distributed



with mean $np$ and standard deviation $\sqrt{np(1-p)}$ **and** we should use the continuity correction (helps when approximating the discrete distribution of $Y$ using the continuous normal distribution).