Section 7.3 Sampling Distribution of the Sample Proportion

We have a population of interest and want to know the proportion, p, of people in favor of something.

How to estimate p?

We take a ______________________

and count X=number of people in favor. We can estimate p as _______________

In repeated samples of the same size n, the values of X will _______________

The distribution of X is ______________________

Normal Approximation for a Count

When X is a count having the $B(n, p)$ distribution and n is large, then

X is approximately

Sampling Distribution for Proportions

The count X of the number of successes is directly related to the proportion of successes $\widehat{P} = \frac{X}{n}$.

Mean of $\widehat{P}$

That is, we would expect the value of a sample proportion to be equal to the population proportion, on average. This says that $\widehat{P}$ is an _______________ estimator for p.

Standard Deviation of $\widehat{P}$

That is, as n increases the standard deviation decreases, which is good, improves the accuracy of the estimator overall.

Since we do not know the value of p, we cannot determine the s.d. of $\widehat{P}$. We can replace the true proportion p

by the sample proportion $\widehat{p}$. The resulting value is called the **standard error of the sample proportion**.

We could convert any probability question about a proportion to that of a count, and vice-versa.

If **n is small**, we would need to convert the question to a count and use the binomial distribution to work it out. If **n is large**, we could convert the question to a count and use the normal approximation for a count, OR use a related normal approximation for a proportion (for large n).

Normal Approximation for a Proportion

Let $\widehat{P}$ be the sample proportion of successes in a random sample of size n

from a population with proportion of successes p.

If n is large, then $\widehat{P}$ is approximately

(rule: $np \geq 10$ and $n\,(1-p) \geq 10$)

Example: According to a market research firm, 52% of all residential telephone numbers in Los Angeles are unlisted. A telemarketing company uses random digit dialing equipment that dials residential numbers at random, regardless of whether they are listed in the telephone directory. The firm calls 500 numbers in Los Angeles.

a. What is the exact distribution of X = the number of unlisted numbers that are called?

b. Compute the probability that at least half of the numbers called are unlisted.

Section 7.4 Estimates That Are Approximately Normal

- Unknown **parameter**: θ
- **estimate**: $\widehat{\theta}$

Bias, precision and accuracy

The **bias** in an estimator is the difference between $\widehat{\theta}$ and θ. A smaller bias is preferred to a larger one.

An estimate $\widehat{\theta}$ is **unbiased** if _______________

The **precision** of an estimate refers to its **variability** in repeated sampling.

One estimate is less precise than another if it has more variability.

Less variability is preferred.

Pictures:

The **standard error** of an estimate $\widehat{\theta}$ [denoted se($\widehat{\theta}$)]

- estimates the variability of $\widehat{\theta}$ values in repeated sampling and
- is a measure of the precision of $\widehat{\theta}$.

Ideally, an **accurate** estimate should have _______________ bias and _______________ precision.

Mean Squared Error (MSE)

is a good and commonly used measure for **accuracy** of an estimate.

A good estimate should have _______________ MSE.

Section 7.5 Standard Errors of Differences

Recall that for two **independent** r.v. X_1 and X_2,

$\operatorname{var}(X_1 - X_2) =$

$\operatorname{sd}(X_1 - X_2) =$

Now, if $\widehat{\theta}_1$ and $\widehat{\theta}_2$ are two independent estimates (e.g., from two independent samples), then

$\operatorname{sd}(\widehat{\theta}_1 - \widehat{\theta}_2) =$

The **Standard Error for a difference** between two independent estimates is

$\operatorname{se}(\widehat{\theta}_1 - \widehat{\theta}_2) =$

The **Standard Error for a difference** between two proportions (from independent samples)

The **Standard Error for a difference** between two means (from independent samples)

Example: Does breast feeding increase the IQ of babies? Data were collected from 5 special-care baby units in England. The babies were all preterm or had low birth weights. Mothers chose whether to provide breast milk for their infant within 72 hours of delivery. The children were given an intelligence test after 8 years of follow-up. Here are the data summary.

Group	**sample size**	**sample mean**	**sample standard deviation**
Bottle fed	90	92.8	15.18
Breast fed	210	103.0	17.39

a. How many standard errors separate the sample means? Are you fairly certain that there is a difference between the true means?

b. Calculate a two-standard-error interval for the difference in true means.

c. What populations do these results apply to?

d. Do the results prove that breast feeding increase IQ? Why or why not?

Section 7.6 Student's t-Distribution

Scenario 1:

We wanted to estimate or test hypotheses about the population mean μ. We assumed we had a random sample from a normal population with **knwon** σ. Our confidence intervals and tests will be based on the following quantity:

$$Z = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}}$$

Notes:

1. The denominator $\frac{\sigma}{\sqrt{n}}$ is

2. The Z statistic has a

Scenario 2:

We still want to estimate or test hypotheses about the population mean μ. We still have a random sample from a normal population but with **unknwon** σ. Our confidence intervals and tests will be based on the following quantity:

$$T = \frac{\overline{X} - \mu}{S/\sqrt{n}}$$

Notes:

1. The denominator $\frac{S}{\sqrt{n}}$ is

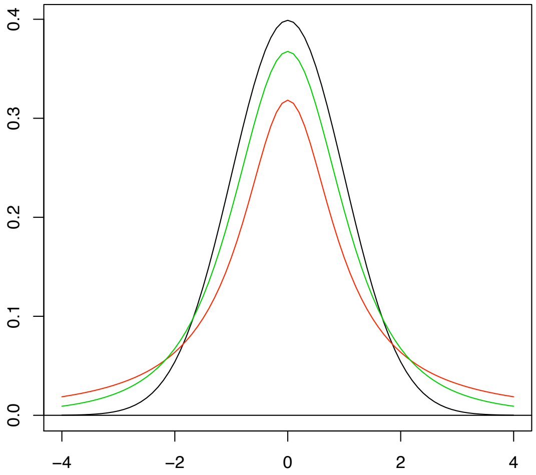
2. The T statistic has a

The T statistic will be used in statistical inferences, confidence intervals and testing hypotheses in Chapters 8 and 9.

About the t-distributions ...

- Symmetric, unimodal, centered at 0
- Flatter with heavier tails compared to the $N(0, 1)$
- As n increases, $t(n)$ ______________________
- The standard normal distribution $N(0, 1)$ is a t-distribution with _______________
- Appendix A6 summarizes percentiles for various t-distributions

Here are the density curves for the standard normal $N(0, 1)$, $t(1)$ and $t(3)$ distributions.



Example – Using the t table

(a) Find 90% percentile for n=12 observations.

(b) Find t^* such that the central region has 95% prob. for n=30 observations.