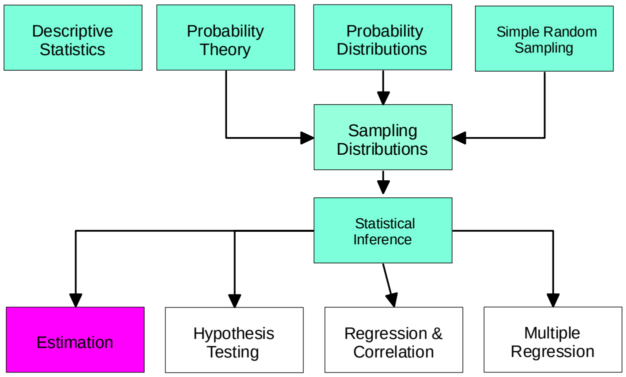
The Course So Far:



- **Estimators:** Random variables used to estimate population parameters.
- **Estimates:** Specific values of an estimator.

$\bar{x}$ is an estimator of μ

s^2 is an estimator of σ^2

$\bar{x} = 120$ is a specific estimate of μ

$s^2 = 237.1$ is a specific estimate of σ^2

- **Point Estimates:** Specify a single value of a population parameter.
- **Interval Estimates:** Specify a range of value of estimates.

$115 \le \bar{x} \le 125$ is an interval estimate of μ

Desirable Properties of Estimators:

✓ **Unbiasedness:** An estimator is said to be unbiased if the expected value of the estimator is equal to the parameter being estimated, or

$$E\left[\hat{\Theta}\right] = \Theta.$$

Proof:

$$E[x/n] = \frac{1}{n}E[x] = \frac{1}{n}(np) = p.$$

✓ **Efficiency:** The most efficient estimator among a group of unbiased estimators is the one with the smallest variance.

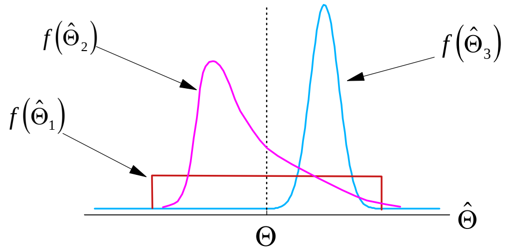
The property of efficiency of an estimator is defined by comparing its variance to the variance of all other unbiased estimators.

Three different estimators based on sample size *n*

$f(\hat{\Theta}_1)$ is unbiased

$f(\hat{\Theta}_2)$ is unbiased

$f(\hat{\Theta}_3)$ is biased



$$V\left[\hat{\Theta}_1\right] > V\left[\hat{\Theta}_2\right] > V\left[\hat{\Theta}_3\right]$$

But estimator 3 is biased, so in this group, estimator 2 is most efficient.

Desirable Properties of Estimators (concluded):

✓ **Sufficiency:** an estimator is sufficient if it uses all the information about the population parameter that the sample can provide**.**

✓ **Consistency:** An estimator is said to be consistent if it yields estimates which approach the population parameter being estimated as *n* becomes larger.

Property of Consistency: $V\left[\hat{\Theta}\right] \to 0$ as $n \to \infty$

bias in $\hat{\Theta}$ $\to 0$ as $n \to \infty$

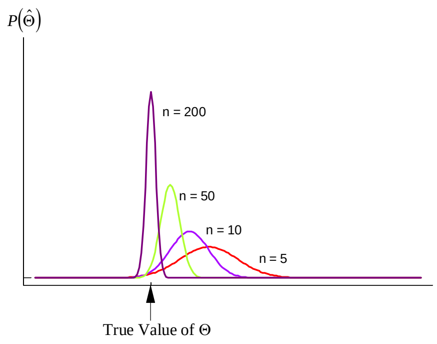
Proof that the sample proportion *x/n* is a consistent estimator of *p*, the population proportion:

$$V[x/n] = \frac{1}{n^2}V[x]$$

$$= \frac{1}{n^2}(npq)$$

$$= \frac{pq}{n} \Rightarrow \lim_{n\to\infty}\left(\frac{pq}{n}\right) = 0$$

A consistent estimator of Θcollapses around Θas sample size increases:



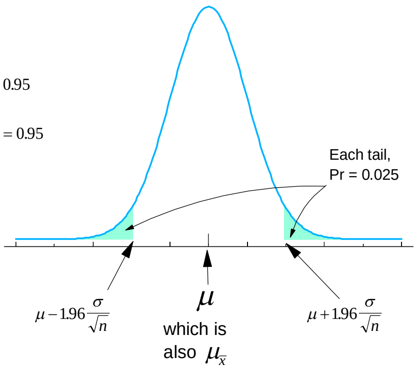
Note: $\hat{\Theta}$ can be biased but still be consistent.

Sampling distribution of the sample mean

$$\Pr\left(-1.96 < \frac{\bar{x}-\mu}{\sigma/\sqrt{n}} < 1.96\right) = 0.95$$

$$\Pr\left(-1.96\frac{\sigma}{\sqrt{n}} < \bar{x}-\mu < 1.96\frac{\sigma}{\sqrt{n}}\right) = 0.95$$

$$\Pr\left(\mu-1.96\frac{\sigma}{\sqrt{n}} < \bar{x} < \mu+1.96\frac{\sigma}{\sqrt{n}}\right) = 0.95$$


Each tail,
Pr = 0.025
$\mu-1.96\frac{\sigma}{\sqrt{n}}$
μ
which is
also $\mu_{\bar{x}}$
$\mu+1.96\frac{\sigma}{\sqrt{n}}$

Sampling distribution of the sample mean

$$\Pr\left(-1.96 < \frac{\bar{x}-\mu}{\sigma/\sqrt{n}} < 1.96\right) = 0.95$$

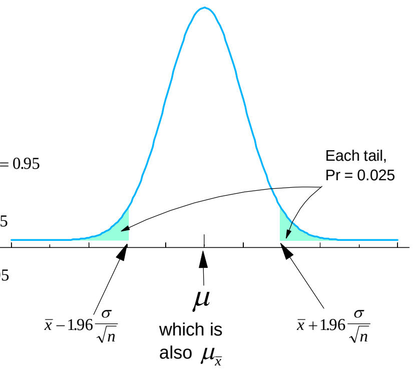
solving for μ:

$$\Pr\left(-1.96\frac{\sigma}{\sqrt{n}} < \bar{x}-\mu < 1.96\frac{\sigma}{\sqrt{n}}\right) = 0.95$$

$$\Pr\left(-\bar{x}-1.96\frac{\sigma}{\sqrt{n}} < -\mu < -\bar{x}+1.96\frac{\sigma}{\sqrt{n}}\right) = 0.95$$

$$\Pr\left(\bar{x}+1.96\frac{\sigma}{\sqrt{n}} > \mu > \bar{x}-1.96\frac{\sigma}{\sqrt{n}}\right) = 0.95$$

$$\Pr\left(\bar{x}-1.96\frac{\sigma}{\sqrt{n}} < \mu < \bar{x}+1.96\frac{\sigma}{\sqrt{n}}\right) = 0.95$$


Each tail,
Pr = 0.025
$\bar{x}-1.96\frac{\sigma}{\sqrt{n}}$
μ
which is
also $\mu_{\bar{x}}$
$\bar{x}+1.96\frac{\sigma}{\sqrt{n}}$

First, let's construct a confidence interval around $\bar{x}$ using our equation:

$$\Pr\left(\mu-1.96\frac{\sigma}{\sqrt{n}} < \bar{x} < \mu+1.96\frac{\sigma}{\sqrt{n}}\right) = 0.95$$

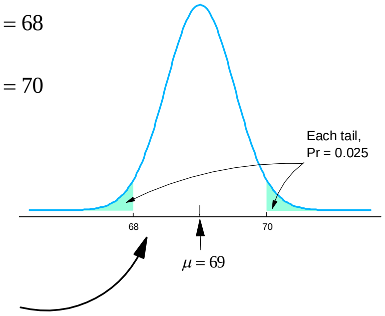
where, by assumption: $\mu = 69$, $\sigma = 3.2$, $n = 36$

$$\mu-1.96\frac{\sigma}{\sqrt{n}} = 69-1.96\frac{3.2}{\sqrt{36}} = 68$$

and,

$$\mu+1.96\frac{\sigma}{\sqrt{n}} = 69+1.96\frac{3.2}{\sqrt{36}} = 70$$

So, assuming we know the true population mean and standard deviation, the 95% confidence interval for sample means computed from samples of size 36 is (68-70) inclusive.



That is, on average 95/100 samples will have $\bar{x}$ in this interval.

$$\Pr\left(\bar{x}-1.96\frac{\sigma}{\sqrt{n}} < \mu < \bar{x}+1.96\frac{\sigma}{\sqrt{n}}\right) = 0.95$$

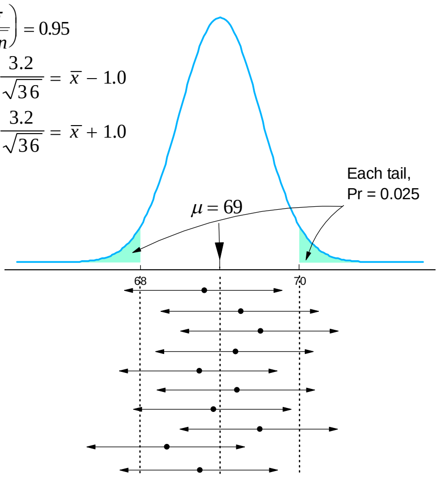
$$\bar{x} - 1.96\frac{\sigma}{\sqrt{n}} = \bar{x} - 1.96\frac{3.2}{\sqrt{36}} = \bar{x} - 1.0$$

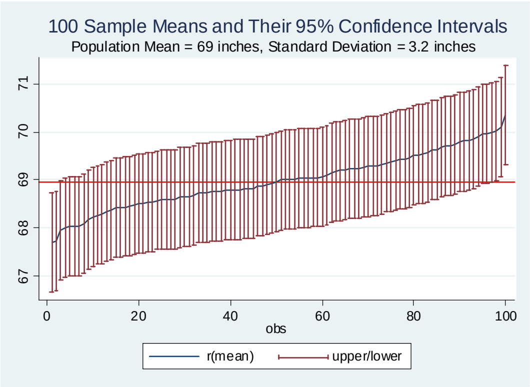
and,

$$\bar{x} + 1.96\frac{\sigma}{\sqrt{n}} = \bar{x} + 1.96\frac{3.2}{\sqrt{36}} = \bar{x} + 1.0$$

Sample Means, 10 samples
n=36

68.7931	69.3153	68.7153
69.2706	68.9306	68.8467
69.4822	69.185	
69.4069	68.2681	





$$s = \sqrt{\frac{1}{n-1}\sum\left(x_i-\bar{x}\right)^2}$$

$$\bar{x} \sim t\left(\mu, \frac{s}{\sqrt{n}}\right)$$

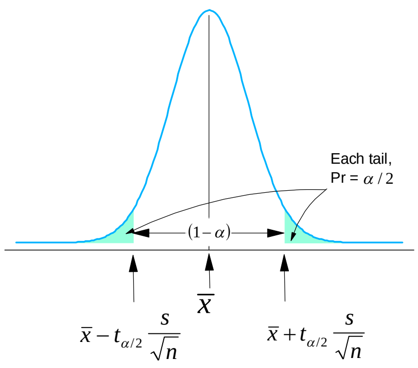
$$t = \frac{\bar{x}-\mu}{s/\sqrt{n}}$$

$$\Pr\left(-t_{\alpha/2} < \frac{\bar{x}-\mu}{s/\sqrt{n}} < t_{\alpha/2}\right) = 1-\alpha$$

$$\Pr\left(\bar{x}-t_{\alpha/2}\frac{s}{\sqrt{n}} < \mu < \bar{x}+t_{\alpha/2}\frac{s}{\sqrt{n}}\right) = 1-\alpha$$

$\bar{x} \pm t_{\alpha/2}\frac{s}{\sqrt{n}}$ The interval estimator of the sample mean

Confidence interval for the sample mean



Each tail,
Pr = $\alpha/2$
$(1-\alpha)$
$\bar{x}$
$\bar{x}-t_{\alpha/2}\dfrac{s}{\sqrt{n}}$
$\bar{x}+t_{\alpha/2}\dfrac{s}{\sqrt{n}}$

Finding a t-interval for a population mean when sigma is unknown:

Assumptions:

1. Normal population or large sample
2. sigma unknown

Step1: For a confidence level of $1-\alpha$ use a t-distribution table to find $t_{\alpha/2}$ with *df = n-1,* where *n* is the sample size.

Step 2: The confidence interval for μ is from

$$\bar{x}-t_{\alpha/2}\frac{s}{\sqrt{n}} \text{ to } \bar{x}+t_{\alpha/2}\frac{s}{\sqrt{n}}$$

where $t_{\alpha/2}$ is found in Step 1 and $\bar{x}$ and *s* are computed from the sample data.

Step 3: Interpret the confidence interval.

The confidence interval is exact for normal populations and is approximately correct for large samples from nonnormal populations.

Remember also, as sample size gets larger the t-distribution approaches the normal distribution and we can use the normal distribution even if we don't know sigma and have to estimate it.

Pickpocket Losses ($ per incident): FBI Data
Random Sample of 25 observations

loss
26
207
214
217
223
253
277
313
397
430
443
447
549
554
570
587
627
649
653
768
805
833
844
883
1064

Are the data normally distributed? Are there outliers to worry about?

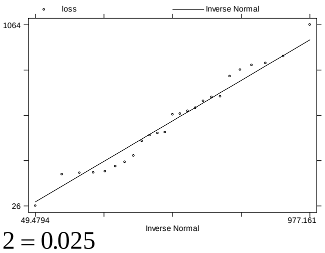
Check normality using the *qnorm procedure:*

```
. summarize loss, detail

                            loss
-------------------------------------------------------------
      Percentiles      Smallest
 1%           26             26
 5%          207            207
10%          214            214       Obs                  25
25%          277            217       Sum of Wgt.          25

50%          549                      Mean             513.32
                        Largest       Std. Dev.      262.2309
75%          653            833
90%          844            844       Variance       68765.06
95%          883            883       Skewness       .1689009
99%         1064           1064       Kurtosis       2.230908
```

```
qnorm loss
```



$$\bar{x}=513.32$$

$$s=262.2309$$

$$1-\alpha=0.95 \text{ so } \alpha=0.05 \text{ and } \alpha/2=0.025$$

$$\bar{x}-t_{\alpha/2}\frac{s}{\sqrt{n}} \text{ to } \bar{x}+t_{\alpha/2}\frac{s}{\sqrt{n}} \qquad t_{0.05/2}=t_{0.025}=2.064 \text{ for } n-1=24 \text{ degrees of freedom}$$

$$513.32-2.064\bullet\frac{262.23}{\sqrt{25}} \text{ to } 513.32+2.064\bullet\frac{262.23}{\sqrt{25}}$$ or 405.07 to 621.57

We can be 95% confident that the mean value lost, μ, of all the year's pickpocket offenses is somewhere between $405.07 and $621.57.

Confidence intervals for Sample Proportions:

$$E[x/n]=\frac{1}{n}E[x]=\frac{1}{n}(np)=p.$$

The sample proportion is an unbiased estimator of the population proportion

The variance of the sample proportion is:

$$V\left[\frac{x}{n}\right]=\frac{1}{n^2}V[x]=\frac{1}{n^2}npq=\frac{pq}{n}, \text{ or}$$

$$\sigma_{\frac{x}{n}}=\sqrt{\frac{pq}{n}}$$

Probability statement for a sample mean:

$$\Pr\left(-1.96<\frac{\bar{x}-\mu}{\sigma_{\bar{x}}}<1.96\right)=\Pr\left(-1.96<\frac{\bar{x}-\mu}{\frac{\sigma}{\sqrt{n}}}<1.96\right)=0.95$$

Probability Statement for a Sample Proportion

$$\Pr\left(-1.96<\frac{\hat{p}-p}{\sigma_{\frac{x}{n}}}<1.96\right)=\Pr\left(-1.96<\frac{\hat{p}-p}{\sqrt{\frac{pq}{n}}}<1.96\right)=0.95$$

Confidence Interval for a Sample Proportion:

$$\Pr\left(-\hat{p}-1.96\sqrt{\frac{pq}{n}}<-p<-\hat{p}+1.96\sqrt{\frac{pq}{n}}\right)=0.95$$

$$\Pr\left(\hat{p}-1.96\sqrt{\frac{pq}{n}}<p<\hat{p}+1.96\sqrt{\frac{pq}{n}}\right)=0.95$$ or, more generally

$$\Pr\left(\hat{p}-z_{\alpha/2}\sqrt{\frac{pq}{n}}<p<\hat{p}+z_{\alpha/2}\sqrt{\frac{pq}{n}}\right)=1-\alpha$$

So, a $1-\alpha$ confidence interval for *p* is:

$$\hat{p}\pm z_{\alpha/2}\sqrt{\frac{pq}{n}}$$

Remember, *n* must be sufficiently large so that the sampling distribution of $\hat{p}$ can be approximated by a normal distribution.

$$\hat{p}\pm z_{\alpha/2}\sqrt{\frac{\hat{p}\hat{q}}{n}}$$

Example: proportion of families in a population that own two cars:

$$n=144$$

$$x=48$$

$$\hat{p}=\frac{48}{144}=\frac{1}{3}=0.333$$

$$s=\sqrt{\frac{\hat{p}\hat{q}}{n}}=\sqrt{\frac{\left(\frac{1}{3}\right)\left(\frac{2}{3}\right)}{144}}=0.0393$$

Assuming a 95% confidence interval:

$$\hat{p}\pm z_{0.05/2}\sqrt{\frac{\hat{p}\hat{q}}{n}}=0.333\pm1.96(0.0393)=0.333\pm0.077$$

$$\text{or } 0.333-0.077<p<0.333+0.077 \Rightarrow 0.256<p<0.410$$

We can conclude with 95% confidence that the population proportion of families who own two or more cars is between 25.6 and 41.0 percent.

Defective Batteries:

Buy 10,000 batteries
Sample 300 batteries
42 are defective

Can we be 99% sure that more than 10% are defective in the purchase of 10,000?

Set up a 99% confidence interval around 42/300 = 0.14 assuming the sampling distribution is normal (central limit theorem)

$$\hat{p} \pm z_{\alpha/2}\sqrt{\frac{\hat{p}\hat{q}}{n}} \Rightarrow 0.14 \pm 2.575\sqrt{\frac{0.86\times 0.14}{300}} = 0.14 \pm 2.575\sqrt{.000401} = 0.14 \pm 2.575\times .02003 = 0.14 \pm 0.0516$$

$$0.0884 \le p \le 0.192$$

The 99% confidence interval contains 0.10, so we can't be 99% confident that more than 10% of the batteries are defective.

However, looking in a normal table, we can find a value for $z_{\alpha/2}$ such that :

$$\hat{p} - z_{\alpha/2}\sqrt{\frac{\hat{p}\hat{q}}{n}} \ge 0.10$$

That value is $z_{\alpha/2} = 2.0$

which corresponds to a confidence interval of 0.9544.

So, adding an additional 0.0228 (right tail)we can be 97.72% sure that at least 10% of the batteries are defective.

Questions to ask before determining optimal *n:*

- What level of confidence do you want to have (i.e., the value of *100(1- α)*?
- What is the *maximum* difference *(D)* you want to permit between the estimate of the population parameter, $\hat{\Theta}$, and the true population parameter, Θ .

(i.e., the value of *D* is the amount of "error" you want to allow in estimating Θ, where $\left|\hat{\Theta} - \Theta\right| \le D)$?

$$z = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}} \sim N(0,1)$$

$$\Pr\left(-z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}} < \bar{x} - \mu < +z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha$$

$$-z_{\alpha/2}\frac{\sigma}{\sqrt{n}} < \bar{x} - \mu < z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

$$\bar{x} - \mu < z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

$$z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

$$D = z_{\alpha/2}\frac{\sigma}{\sqrt{n}}.$$

Solving for *n:*

$$\frac{D}{z_{\alpha/2}\sigma} = \frac{1}{\sqrt{n}}$$

$$\sqrt{n} = \frac{z_{\alpha/2}\sigma}{D}$$

$$n = \frac{z_{\alpha/2}^2\sigma^2}{D^2}.$$

The optimal sample size.

$$t = \frac{\bar{x} - \mu}{s / \sqrt{n}} \sim t(0,1)$$

Finding optimal sample size for a population proportion:

Choose a maximum difference between estimate and true value of population proportion:

$$\left|\hat{p} - p\right| = D$$

$$\sqrt{\frac{1}{4n}}$$

This is the maximum value for the population standard deviation

$$z = \frac{\hat{p} - p}{\sqrt{\frac{pq}{n}}} \Rightarrow z = \frac{\hat{p} - p}{\sqrt{\frac{1}{4n}}}$$

$$-z_{\alpha/2}\sqrt{1/4n} \le \hat{p} - p \le z_{\alpha/2}\sqrt{1/4n}$$

$$D = z_{\alpha/2}\sqrt{1/4n}$$

and, solving for *n:*

$$D^2 = z_{\alpha/2}^2\frac{1}{4n}$$

$$n = \frac{z_{\alpha/2}^2}{4D^2}$$

The optimal sample size for a sample proportion.

This is conservative, since we've used the largest possible population standard deviation