## Example 69: Opinion Polling

Suppose that 40 % of the population support a certain political candidate.

Given a random sample of 150 individuals, find

1.) the expected value and variance of the number of sampled individuals that favour the candidate.

2.) the probability that more than half the sample favour the candidate.

## Example 69: Opinion Polling

Suppose that 40 % of the population support a certain political candidate.

Given a random sample of 150 individuals, find

1.) the expected value and variance of the number of sampled individuals that favour the candidate.

2.) the probability that more than half the sample favour the candidate.

Let $X_i$ be the answer of the i-th person, "yes" meaning 1, and "no" meaning 0.

$\Rightarrow X_i \sim \text{Bernoulli}(p)$ with $p = 0.4$

Let $Y := \sum_{i=1}^{n} X_i \Rightarrow Y \sim \text{Binom}(n, p)$, with $n = 150$

$\Rightarrow$ $E[Y] = n \cdot p = 150 \times 0.4 = 60$

$Var(Y) = n \cdot p \cdot (1-p) = 150 \times 0.4 \times 0.6 = 36$

Check the rule of thumb:

$n \times p = 60 > 5, \quad n \times (1-p) = 90 > 5$

$\Rightarrow$ Approximate $Y$ by $N(60, 36)$.

We want

$P[Y > 75]$

How can we compute this?

1) <u>Use the Binomial:</u>

Let $\varphi$ be the cdf of Binom(150, 0.4).

R delivers

$$P[Y > 75] = 1 - P[Y \leq 75] = 1 - \varphi(75)$$

$$= 0.005225$$

2) <u>Approximate $Y$ by a $Y' \sim N(60, 36)$</u>

$$P[Y > 75] \approx P[Y' > 75.5]^{*} = 1 - \Phi_{60,36}(75.5)$$

$$= 0.004892 \qquad \text{(with R)}$$

$$1 - \Phi\left(\frac{75.5 - 60}{6}\right)$$

* continuity correction

## 3) Approximation and Lookup in Z-Table

Transform $Y'$ to $Z \sim \mathcal{N}(0,1)$ :

$$P[Y' > 75.5] = P\left[\frac{Y'-60}{6} > \frac{75.5-60}{6}\right]$$

$$\approx P\left[Z > \frac{75.5-60}{6}\right] = P\left[Z > \frac{15.5}{6}\right]$$

$$= P[Z > 2.583] = 1 - \Phi(2.583)$$

$$\approx 1 - 0.9951 = 0.0049$$

# How Many Measurements are Needed?

We can use the CLT to determine the number of measurements needed for a required accuracy if we know the variance of the distribution of measurements.

Example 66: We want to measure the distance to a star with

- accuracy $a = 1$ (i.e., with absolute error $\le \frac{a}{2} = 0.5$) and
- certainty $\gamma = 95\%$.

The variance of the measurements is $\sigma^2 = 2^2$.

Let $d$ be the exact distance and $X_i$ be the measurements.
The sample mean $\overline{X}_n$ is close to a normal with

$$\mu_n = \mu \quad \text{and} \quad \sigma_n^2 = \frac{\sigma^2}{n}.$$

Then

$$\frac{\overline{X}_n - \mu_n}{\sigma_n} = \frac{\overline{X}_n - \mu}{\sigma / \sqrt{n}} \sim N(0,1) \text{ approximately.}$$

We want $n$ such that

$$P\left[-\frac{a}{2} < \bar{X}_n - \mu < \frac{a}{2}\right] \leq \gamma$$

That is

$$\gamma \leq P\left[-\frac{\sqrt{n}}{\sigma}\frac{a}{2} < \frac{\sqrt{n}}{\sigma}(\bar{X}_n - \mu) < \frac{\sqrt{n}}{\sigma}\frac{a}{2}\right]$$

$$\approx P\left[-\frac{\sqrt{n}}{\sigma}\frac{a}{2} < Z < \frac{\sqrt{n}}{\sigma}\frac{a}{2}\right]$$

$$= 1 - 2\left(1 - \Phi\left(\sqrt{n}\frac{a}{2\sigma}\right)\right) = 2\cdot\Phi\left(\sqrt{n}\frac{a}{2\sigma}\right) - 1,$$

hence

$$\Phi\left(\sqrt{n}\frac{a}{2\sigma}\right) \geq \frac{1+\gamma}{2}$$

$$\iff \sqrt{n}\frac{a}{2\sigma} \geq \Phi^{-1}\left(\frac{1+\gamma}{2}\right)$$

$$\iff \sqrt{n} \geq \frac{2\sigma}{a}\Phi^{-1}\left(\frac{1+\gamma}{2}\right)$$

This is an example where we need the inverse of the cdf to reason backward from a probability to an argument.

We need an $n$ such that

$$\sqrt{n} \geq \frac{2\sigma}{a} \Phi^{-1}\left(\frac{1+\gamma}{2}\right)$$

with

$$a = 1, \quad \sigma = 2, \quad \gamma = 0.95.$$

This yields

$$\sqrt{n} \geq \frac{2 \cdot 2}{1} \Phi^{-1}\left(\frac{1+0.95}{2}\right) = 4 \cdot \Phi^{-1}(0.975)$$

$$= 4 \times 1.960 \quad (\text{in Z-table})$$

$$= 4 \times 1.959964 \quad (\text{with R})$$

Hence $n \geq (4 \times 1.960)^2 = 61.4656$

is a sufficiently large number of measurements

## 4.2 Sample Variance

If we make measurements of some quantity, we consider this as evaluating a RV $X$. If we make several measurements, then we consider them as evaluations of $n$ RVs $X_1, \ldots, X_n$ that are i.i.d., having the same distribution as $X$.

How can we estimate the mean value of the distribution of $X$, i.e., $E[X]$?

The average $\overline{X}_n$ of the $X_i$, $\overline{X}_n = \frac{1}{n} \sum_{i=1}^{n} X_i$, should be a good estimate.

How can we check that this is conceptually the right thing to do?

# Unbiased Estimators

Suppose $X, X_1, \dots, X_i, \dots$ are i.i.d. RVs.

A function $F(x_1, \dots, x_n)$, if applied to $X_1, \dots, X_n$, defines a new random variable $F(X_1, \dots, X_n)$.

An example is $\overline{X}_n$, which is defined by

$$F(x_1, \dots, x_n) = \frac{1}{n}(x_1 + \dots + x_n) = \overline{x}_n.$$

<u>Definition</u>: Let $X_1, \dots, X_n$ be i.i.d. RVs, $F: \mathbb{R}^n \to \mathbb{R}$ a function and $\theta$ be a parameter (like mean, variance, or skew) of the distribution of the $X_i$.

Then the bias of $F$ with respect to $\theta$ for $X_1, \dots, X_n$ is

$$E(F(X_1, \dots, X_n)) - \theta,$$

and $F(X_1, \dots, X_n)$ is an unbiased estimator if the bias is $0$.

Examples: (1) The average $\overline{X}_n = \frac{1}{n} \sum_{i=1}^{n} X_i$ is an unbiased estimator of the mean $\mu$.

(2) The average squared distance from the mean

$$\frac{1}{n} \sum_{i=1}^{n} (X_i - \mu)^2$$

is an unbiased estimator of the variance. (Note that we used $\mu$, not $\overline{X}_n$.)

Proof: (1) We have calculated several times that

$$E[\overline{X}_n] = \frac{1}{n} \sum_{i=1}^{n} E[X_i] = \frac{1}{n} \cdot n \cdot \mu = \mu$$

(2) Remember that $Var(X) = E[(X-\mu)^2]$. Thus

$$E\left(\frac{1}{n} \sum_{i=1}^{n} (X_i - \mu)^2\right) = \frac{1}{n} \sum_{i=1}^{n} E(X_i - \mu)^2 = \frac{1}{n} \cdot n \cdot \sigma^2 = \sigma^2$$

# Estimating the Variance

Consider the function

$$T^2(x_1, \dots, x_n) = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2 \quad {}^{*)}$$

with $\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$.

Then one can calculate (see lecture notes of 19/20) that

$$E[\, T^2(X_1, \dots, X_n)] = \frac{n-1}{n} \, Var(X).$$

Thus, this is an estimator with bias! But

$$\frac{n}{n-1} T^2(x_1, \dots, x_n) = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2 =: S^2$$

is unbiased! This is also called the sample variance.

---

*) "$T^2$" is an abuse of notation, motivated by the attempt to estimate the variance